Extraction de relations spatio-temporelles à partir des données environnementales et de la santé

Hugo Alatrista-Salas

ACADÉMIE DE MONTPELLIER
UNIVERSITÉ **M**ONTPELLIER **II**
Sciences et **T**echniques du **L**anguedoc

THÈSE

présentée au Laboratoire d'Informatique de Robotique
et de Microélectronique de Montpellier
et à l'Université de Nouvelle Calédonie pour
obtenir le diplôme de doctorat

Spécialité	:	**Informatique**
Formation Doctorale	:	**Informatique**
École Doctorale	:	**Information, Structures, Systèmes**

Extraction de relations spatio-temporelles à partir des données environnementales et de la santé

par

Hugo Alatrista-Salas

Soutenue le 04 octobre 2013, devant le jury composé de :

Directrices de thèse
Mme. Maguelonne TEISSEIRE, directrice de recherche ..... TETIS, Irstea, FRANCE
Mme. Nazha SELMAOUI-FOLCHER, maître de conference, HDR ..... Université de NOUVELLE CALEDONIE

Rapporteurs
Mme. Florence LE BER, ingénieur en chef PEF, HDR ..... Ecole Nationale du Génie de l'Eau et de l'Environnement de Strasbourg, FRANCE
M. Osmar ZAÏANE, professeur ..... University of Alberta, CANADA

Examinateurs
Mme. Karine ZEITOUNI, professeur ..... Université Versailles, FRANCE
M. Frédéric FLOUVAT, maître de conference ..... Université de NOUVELLE CALEDONIE
Mme. Sandra BRINGAY, maître de conference ..... Université Montpellier III, FRANCE
M. Jérôme AZÉ, professeur ..... Université Montpellier 2, FRANCE

Ignorance is the curse of God,
knowledge the wing wherewith we
fly to heaven.

W. SHAEKSPEARE

Remerciements

Ce manuscrit conclut trois ans de travail, je tiens en ces quelques lignes à exprimer ma reconnaissance envers tous ceux qui de près ou de loin y ont contribué.

Incontestablement cette thèse n'aurait jamais pu être menée à son terme sans un encadrement humain et scientifique. Donc, je tiens à remercier de tout mon cœur Mme. Maguelonne Teisseire, Mme. Nazha Selmaoui-Folcher, Mme. Sandra Bringay et M. Frédéric Flouvat. Merci de m'avoir permis de réaliser cette thèse sous vos regards, merci pour votre aide précieuse sur le plan scientifique, administratif et aussi personnel. Merci pour votre patience et votre disponibilité. Merci à vous tous pour ces années qui m'ont appris que la recherche est un métier dont les fruits sont toujours doux. Sans aucun doute, vous êtes l'équipe de travail dont tout le monde rêve.

Je tiens à adresser mes plus sincères remerciements à Mme. Florence Le Ber et à M. Osmar Zaïane pour avoir accepté d'être rapporteurs de ce mémoire et pour le temps précieux qu'ils ont consacré à cette tache. Je remercie également Mme. Karine Zeitouni et M. Jérôme Azé d'avoir participé à mon jury de thèse en tant qu'examinateurs.

Ce travail de thèse s'est effectué au sein du laboratoire TETIS, le PPME de l'Université de Nouvelle Calédonie et l'équipe TATOO du Lirmm. Mes remerciements vont donc s'adresser tout d'abord aux responsables de ces équipes et laboratoires : M. Jean-Philippe Tonneau, Mme. Nazha Selmaoui-Folcher et M. Pascal Poncelet, merci de m'avoir accueilli chaleureusement au sein de vos laboratoires en offrant un cadre de travail matériel et intellectuel favorable.

J'adresse aussi tous mes remerciements à mes collègues de travail dont le soutien a été inestimable et sans faille. Donc, je tiens à remercier à mes compagnons de bureau, Nathalie et Eric, merci pour vos encouragements et les moments de détente autour d'un Haribo.

Mes remerciements les plus chaleureux vont également aux collègues qui m'ont soutenu jusqu'aux derniers jours. Je tiens à manifester toute ma gratitude à Mickaël pour son apport à mes recherches, pour ses conseils et son temps dédié à nos discutions. Un grand merci à Arnaud pour ses conseils concernant la visualisation de l'information et ses relectures de ce manuscrit. Merci également à Pierre pour sa précieuse aide au développement du prototype de visualisation. Je ne peux pas oublier Mathieu et Dino qui m'ont épaulé pendant tout le déroulement de ma thèse. Un grand merci à Hai qui m'a accompagné pendant ces trois années et pour les interminables discutions (en anglais) autour d'un café. Finalement, mes remerciements sincères à Abdoulkader, Lilia, Juan, Fábio et Flavien avec lesquels j'ai eu le plaisir de partager une grande amitié.

Enfin, je remercie mes parents et ma sœur pour leur soutien au cours de ces trois années et sans lesquels je n'en serais pas là aujourd'hui.

Résumé

Face à l'explosion des nouvelles technologies (mobiles, capteurs, etc.), de grandes quantités de données localisées dans l'espace et dans le temps sont désormais disponibles. Les bases de données associées peuvent être qualifiées de bases de données spatio-temporelles car chaque donnée est décrite par une information spatiale (e.g. une ville, un quartier, une rivière, etc.) et temporelle (e.g. la date d'un événement). Cette masse de données souvent hétérogènes et complexes génère ainsi de nouveaux besoins auxquels les méthodes d'extraction de connaissances doivent pouvoir répondre (e.g. suivre des phénomènes dans le temps et l'espace).

De nombreux phénomènes avec des dynamiques complexes sont ainsi associés à des données spatio-temporelles. Par exemple, la dynamique d'une maladie infectieuse peut être décrite par les interactions entre les humains et le vecteur de transmission associé ainsi que par certains mécanismes spatio-temporels qui participent à son évolution. La modification de l'un des composants de ce système peut déclencher des variations dans les interactions entre les composants et finalement, faire évoluer le comportement global du système.

Pour faire face à ces nouveaux enjeux, de nouveaux processus et méthodes doivent être développés afin d'exploiter au mieux l'ensemble des données disponibles. Tel est l'objectif de la fouille de données spatio-temporelles qui correspond à l'ensemble des techniques et méthodes qui permettent d'obtenir des connaissances utiles à partir de gros volumes de données spatio-temporelles. Cette thèse s'inscrit dans le cadre général de la fouille de données spatio-temporelles et l'extraction de motifs séquentiels. Plus précisément, deux méthodes génériques d'extraction de motifs sont proposées. La première permet d'extraire des motifs séquentiels incluant des caractéristiques spatiales. Dans la deuxième, nous proposons un nouveau type de motifs appelé "motifs spatio-séquentiels". Ce type de motifs permet d'étudier l'évolution d'un ensemble d'événements décrivant une zone et son entourage proche.

Ces deux approches ont été testées sur deux jeux de données associées à des phénomènes spatio-temporels : la pollution des rivières en France et le suivi épidémiologique de la dengue en Nouvelle Calédonie. Par ailleurs, deux mesures de qualité ainsi qu'un prototype de visualisation de motifs ont été également proposés pour accompagner les experts dans la sélection des motifs d'intérêts.

Mots clefs : *Fouille de données spatio-temporelles, Information géographique, Recherche de corrélations, Exploration de données, Système de détection épidémiologique.*

Abstract

Recently, thanks to the advanced technologies, (e.g. smartphones, sensors, etc.), large amounts of spatiotemporal data are now available. Commonly, a given spatiotemporal dataset contains a set of rows each of which presents spatial and temporal information of a happened event. The spatial information could be a city, a neighborhood, a river, a GPS location, etc. meanwhile temporal information is the date-time of the concerned event. Knowledge extraction from spatiotemporal data has been studied in many years for understanding the evolution or the spreading of phenomena in both temporal and spatio dimensions. However, there are still many challenging issues we need to deal with.

For instance, the dynamics of an infectious disease can be described as: (1) the interactions between humans; (2) the transmission vector as well as (3) some unrevealed spatiotemporal mechanisms involved in its spreading. In fact, the varying of one of these components can trigger changes the interaction scheme between the components and finally alter the behaviour of the whole system.

In my thesis, I will concern on proposing novel spatiotemporal data mining techniques to capture this phenomenon. More specifically, two generic methods of pattern mining are proposed: (1) the first one enables us to extract sequential patterns including spatial characteristics from the data; and (2) we propose a novel type of patterns called spatio-sequential patterns which are used to express the evolution of a set of events in an area and its near environment. Our proposed approaches were tested on real datasets associated to two spatiotemporal phenomena: the spreading of pollution in rivers in France and the epidemiological monitoring of dengue in New Caledonia. In additional, two qualitative measures and a pattern visualization prototype are also supplied to assist the experts in the selection of meaningful patterns.

Keywords: *Spatiotemporal data mining, Geographic information, Correlations research, Data exploration, Epidemiological detection systems.*

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

L'explosion des dispositifs de collecte de données numériques tels que les capteurs, les GPS, etc. et le développement des technologies de stockage des données ont permis aux entreprises et aux organisations de stocker de très grandes quantités de données, rendant ces dernières difficiles à analyser sans outil automatique. Quatre facteurs ont conduits à ce phénomène : (1) les systèmes de stockage temporaires et permanents sont de moins en moins coûteux ; (2) la vitesse de calcul dans les processeurs a augmenté ; (3) la vitesse et la fiabilité de transmission des données se sont améliorées ; et (4) les systèmes de gestion de bases de données sont de plus en plus puissants. Désormais, nous entendons parler de téraoctets, pétaoctets et très bientôt, des exaoctets.

À titre d'exemple, si nous cherchons le mot anglais *"information"* dans Google, nous obtenons 9 680 000 000 réponses nous guidant vers des sites contenant ce mot. Supposons que nous soyons assez rapides pour consulter une page en moins de cinq secondes, il nous faudra un peu plus de 1 500 ans pour toutes les visiter. Evidement, réaliser cette tâche est impossible. Il y a donc un besoin évident de nouvelles technologies pour non seulement stocker et rechercher des informations, mais également pour les analyser et les interpréter.

Outre ces gros volumes, une autre difficulté est liée à leur hétérogénéité. Actuellement, les organisations possèdent souvent plusieurs systèmes de stockage adaptés à des caractéristiques particulières de leurs données (des fichiers textes, des tableurs, plusieurs bases de données, etc.). Dans cette "pluralité" de données, nous pouvons trouver, des données numériques, textuelles - structurées ou non -, des images, des sons et bien d'autres types de données. L'ensemble de ces systèmes n'est pas directement exploitable. Cette hétérogénéité rend difficile la tâche de recherche d'information. Il ne suffit pas de les homogénéiser mais il faut également exprimer les relations existant entre elles.

Pour faire face à ces problèmes, de nouvelles techniques doivent être pensées pour

aider les humains à transformer et résumer automatiquement ces gros volumes de données hétérogènes en connaissances utiles. Ces nouvelles connaissances permettront une meilleure compréhension des phénomènes qui se produisent dans leur environnement.

1.1 Le rôle des données

Les bases de données jouent un rôle important dans le monde d'aujourd'hui. Elles sont utilisées pour stocker des informations représentant l'état d'une organisation ou d'un phénomène au cours du temps. Cette information "historique" est obtenue grâce à l'interaction et l'ubiquité [1] des systèmes informatiques et/ou non informatiques qui nous entourent, par exemple, des capteurs, des téléphones mobiles, des sondages, etc. On parle alors de base de données temporelles où chaque transaction représente l'état d'un ensemble d'évènements à un moment donné [Snodgrass, 1992].

En plus de leur rôle de "mémoire de stockage", les données sont parfois plus complexes qu'une simple transaction. Elles ne représentent pas seulement le passé et l'état actuel d'une organisation ou d'un phénomène, mais elles peuvent être utilisées pour prédire son avenir. Au cours d'une bonne partie du XXème siècle, les sciences inductives, ont établi formellement ce qui peut être prédit à partir des expériences antérieures. Pour une organisation, connaitre en détail son état actuel est crucial pour obtenir une bonne prédiction, soit logique (e.g. la classification) ou temporelle (e.g. des évènements à venir). Autrement dit, nous ne pouvons pas prédire le comportement futur d'un phénomène sans connaître l'état courant de la situation [Rudwick, 1985]. Prédire l'avenir n'a rien de magique et la plupart de nos actions quotidiennes sont fondées sur des prévisions. Notre utilisation quotidienne des prévisions météorologiques en est un exemple classique.

Contrairement aux décisions personnelles, les décisions collectives ont souvent des conséquences importantes (e.g. économiques, écologiques, sanitaires, etc.). Par exemple, aux États Unis, la Société du Climat utilise chaque jour environ 2,5 millions de mesures météorologiques, 150 millions d'observations du sol et 10 trillions de données permettant la construction des scénarios. Ces informations sont ensuite utilisées pour réduire le risque de perte de récolte. De mauvaises conditions météorologiques peuvent entrainer plus de 90% de perte pour les récoltes [2].

Il existe pour cela de nombreuses méthodes permettant d'analyser, de transformer et d'exploiter ces masses de données. Ces méthodes sont regroupées et assemblées dans un processus appelé *Extraction de Connaissances à partir des Données* (ou *Knowledge Discovery in Databases*, en anglais).

1. Faculté d'être partout à la fois.
2. O'Reilly Strata Conference, Santa Clara, Cal. 2012. http://strataconf.com/strata2012/public/schedule/detail/22511

1.2 Extraction de connaissances à partir des données

L'*extraction de connaissances à partir des données* (*ECD*) est un domaine de recherche très dynamique. Fayyad *et al.* [1996] a présenté ce processus dont l'objectif est de *transformer des données de bas niveau sous d'autres formes plus compactes, plus abstraites ou plus utiles*. Fayyad *et al.* [1996] décrit ce processus comme un ensemble d'étapes interactives et itératives. Ce processus peut être très complexe et les étapes peuvent varier considérablement en fonction de la nature des données et des objectifs de l'application. Dans ces étapes, nous trouvons : la sélection des données, le pré-traitement, la transformation, la fouille de données et le post-traitement et l'interprétation. Pour sélectionner les données, il faut tout d'abord déterminer les sources d'informations qui pourront être utiles. Ces sources d'informations peuvent être structurées (e.g. une base de données transactionnelles) ou non structurées (e.g. un document en langue naturelle). Ensuite, les données sont nettoyées et formatées de façon à pouvoir appliquer une technique de fouille de données. L'algorithme de fouille est choisi selon le type des données et la problématique applicative. Finalement, les motifs extraits sont restitués pour être validés par les experts du domaine. Une fois seulement les motifs validés, on obtient des connaissances. La Figure 1.1 illustre les étapes du processus d'ECD.

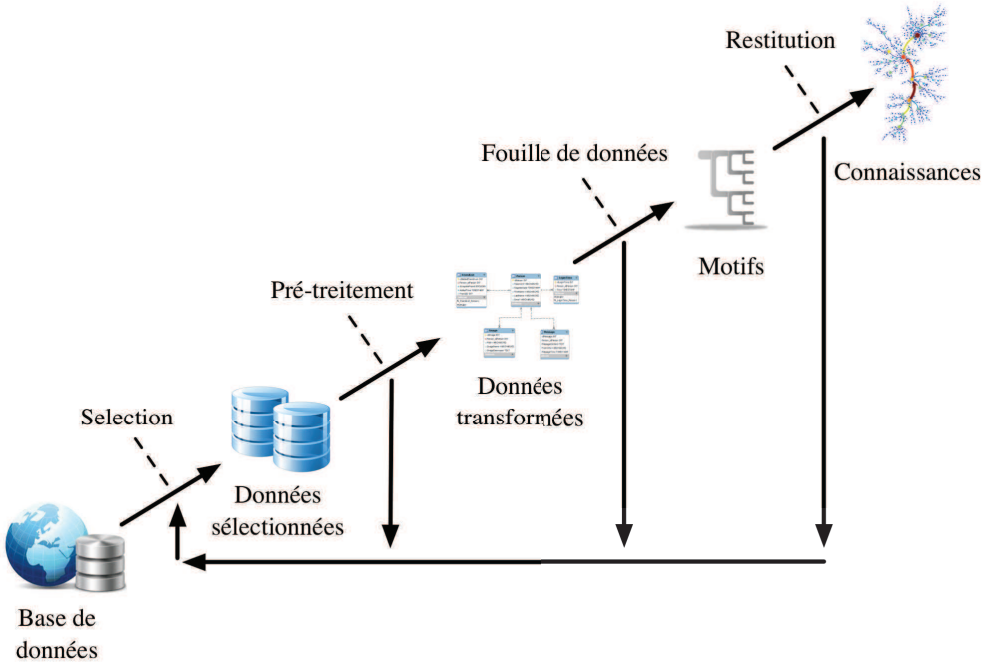


FIGURE 1.1 – Processus d'extraction de connaissances à partir des données (ECD) [Fayyad *et al.*, 1996]

Même si l'étape de fouille de données n'est qu'une partie du processus général d'ECD, elle est celle qui suscite le plus de travaux dans la littérature. La fouille de données est définie comme *le processus non trivial d'extraction d'informations implicites, nouvelles et potentiellement utiles à partir de grands volumes de données* [Fayyad *et al.*, 1996]. Ce domaine connaît une croissance assez spectaculaire, sous l'impulsion des organisations propriétaires de grands volumes de données et soucieuses d'en extraire de la valeur ajoutée. La fouille consiste à induire des lois générales à partir des données stockées. Ces généralisations sont le plus souvent des énoncés de haut niveau, comme par exemple des règles descriptives ou des arbres de décision. La découverte de motifs (*Pattern Discovery*) dans les données est l'un des problèmes phare en fouille de données. Il a été beaucoup étudié en bioinformatique pour l'analyse des données génomiques à large échelle et dans de nombreux domaines d'applications où des régularités peuvent être porteuses de valeur ajoutée (e.g. la découverte de règles d'associations dans des données transactionnelles [Agrawal *et al.*, 1993], de motifs séquentiels dans des bases de séquences de comportements [Agrawal et Srikant, 1995; Mannila *et al.*, 1997; Masseglia *et al.*, 1998] ou découverte de motifs plus complexes tels que des sous-graphes [Inokuchi *et al.*, 2000] ou des sous-arbres [Termier *et al.*, 2002; Zaki, 2002].

Plus récemment, la prise en compte des données spatiales associées à la croissance de l'utilisation des équipements mobiles posent de nouvelles problématiques à la communauté ECD. Ces données sont appelées *données géo-référencées*[3] et permettent de modéliser des phénomènes spatio-temporels (e.g. le suivi d'un orage).

Dans cette thèse, l'objectif principal est de proposer de nouvelles méthodes qui soient adaptées aux données dédiées à l'étude de phénomènes spatio-temporels (i.e., un phénomène évoluant dans l'espace et dans le temps) pour lesquels les caractéristiques spatiales et temporelles des données sont difficiles à intégrer de façon conjointe. Dans la section suivante, nous allons donner la définition d'un phénomène spatio-temporel et sa représentation à l'aide des bases de données.

1.3 Etude de phénomènes spatio-temporels

La plupart des phénomènes observés du monde réel sont de nature dynamique. Yuan [2009] décrit le concept de "dynamique" comme *un ensemble de forces dynamiques influant sur le comportement d'un système et des composants, individuellement et collectivement.* Nous pouvons citer, par exemple, des phénomènes météorologiques comme la pluie, le vent, etc. Ces phénomènes sont généralement associés à des entités spatiales comme des rivières, des villes, etc. (les zones où se déroulent les évènements météorologiques), représentés par des caractéristiques statiques (e.g. le cours de la rivière, la surface

3. L'information géo-référencée est associée à un objet ou à un phénomène qui concerne le monde terrestre en considérant son positionnement sur la surface terrestre, sa nature, son aspect et ses caractéristiques diverses, *Piard, B.E., Centre National de l'Information Géo-Spatiale (C.N.I.G.S).*

d'une ville ou la position d'un capteur) qui ne changent pas dans une période de temps fixe [Asproth *et al.*, 1995]. En outre, des informations dynamiques peuvent être associées aux informations géo-référencées (e.g. la quantité de pluie tombée sur une journée, la force du vent, la température, etc.) et évoluent au cours du temps. Finalement, un phénomène spatio-temporel est un processus lié au changement, parfois périodique, pour lequel les caractéristiques d'un ensemble d'entités spatiales sont exprimées par des séries ou séquences d'évènements [Nadi et Delavar, 2003].

Dans la vie quotidienne, nous pouvons observer de nombreux phénomènes spatio-temporels. Par exemple, le déplacement d'un orage est associé à des informations spatiales (e.g. les coordonnées de départ et d'arrivée) et temporelles (e.g. les dates de début et de fin). La Figure 1.2 représente l'évolution de la météo pour trois zones Z_1, Z_2 et Z_3 à trois dates consécutives.

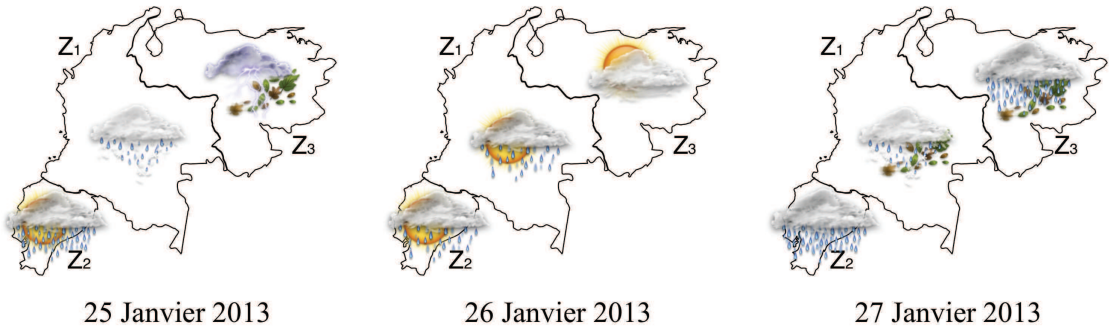


FIGURE 1.2 – Exemple d'un phénomène spatio-temporel : La météo

L'information relative au phénomène illustré dans la Figure 1.2, peut être représentée par une base de données dite "spatio-temporelle". Une base de données spatio-temporelles est un ensemble structuré d'informations incluant des composantes géographiques (e.g. une localisation associée à une granularité particulière comme des villes, des quartiers, des rivières, etc.) et des composantes temporelles (des dates). Une telle base contient également des informations décrivant la composante géographique à un moment donné. Le Tableau 1.1 décrit la base de données associées à l'exemple de la figure précédente.

Ces phénomènes dynamiques et complexes et leurs bases de données spatio-temporelles associées, génèrent de nouveaux besoins auxquels les méthodes d'ECD devraient pouvoir répondre. Souvent, seule la réunion des données spatiales et temporelles permet d'appréhender le problème dans sa globalité. L'objectif de cette thèse est de construire des méthodes automatiques permettant de générer efficacement des modèles spatio-temporels, qui soient capables de capturer l'essence des processus dynamiques complexes comme ils se produisent dans le monde réel.

TABLE 1.1 – Evolution des conditions climatiques pour les zones Z_1, Z_2 et Z_3 pour le 25, 26, 27 Janvier 2013

Zone	Date	Température	Humidité	Precipitations	Rafales
Z_1	25/01/2013	T_b	H_m	P_m	-
Z_1	26/01/2013	T_m	H_m	P_b	-
Z_1	27/01/2013	T_b	H_m	P_m	55
Z_2	25/01/2013	T_m	H_m	P_m	-
Z_2	26/01/2013	T_m	H_m	P_b	-
Z_2	27/01/2013	T_b	H_b	P_m	-
Z_3	25/01/2013	T_b	H_m	P_h	75
Z_3	26/01/2013	T_m	H_h	P_b	-
Z_3	27/01/2013	T_m	H_h	P_h	55

haute (h)
moyenne (m)
basse (b)

Nous motivons et détaillons ces objectifs dans la section suivante pour deux applications liées aux domaines de la santé et l'environnement.

1.4 Applications au suivi de la dengue et de la pollution des rivières

Ces dernières années, l'environnement, le développement durable et la gestion des risques naturels sont devenus des enjeux majeurs, entrainant l'étude de certains phénomènes spatio-temporels ayant une dynamique très complexe. En effet, les avancées technologiques en terme d'acquisition des données (images satellitaires, capteurs, etc.) ont un grand nombre d'applications dans la surveillance et le suivi environnemental tels que la détection de changements abrupts (catastrophes naturelles, etc.), le suivi de phénomènes évolutifs (érosion côtière, désertification, etc.), le suivi de phénomènes qui se propagent (épidémies, pollution de rivières, etc.) ou la mise au point de modèles (hydrologie, activité agricole, etc.) et sont destinées à comprendre et prédire des phénomènes résultant de processus complexes et d'origine pluridisciplinaire (données climatiques, géologiques, médicales, etc.).

Dans cette thèse, nous allons étudier deux phénomènes spatio-temporels issus du domaine de la santé et de l'environnement, particulièrement importants pour les autorités françaises en raison de leurs enjeux économiques :

– les épidémies de dengue en Nouvelle Calédonie : une épidémie désigne l'augmentation rapide de l'incidence d'une maladie contagieuse ou non [4]. Selon la DASS [5], 10 856 cas de dengue en Nouvelle Calédonie ont été enregistrés depuis Septembre

4. Définition proposée pour l'Organisation Mondiale de la Santé (OMS).
5. Direction d'Affaires Sanitaires et Sociales de Nouvelle Calédonie.

2012 dont 1 décès. Il n'existe ni vaccin ni médicament qui protègent contre cette maladie virale qui peut entraîner une fièvre hémorragique parfois mortelle. Mieux comprendre la propagation d'une telle épidémie est donc un enjeu crucial pour la haute autorité de santé afin de déclencher des alertes et mener des campagnes de prévention ;

– la qualité de l'eau des rivières françaises : la croissance rapide de la population due au développement des activités humaines (e.g. l'agriculture, l'industrie, les transports, etc.) a accru la vulnérabilité des ressources hydrographiques. Par exemple, le cumul des déversements réels de nitrates depuis l'année 1971 montre que la baie de Vilaine (France) a dû absorber environ 2 200 000 tonnes de nitrates, soit 500 000 tonnes d'azote [6]. Mieux comprendre l'évolution de ces pollutions est donc un enjeu crucial pour les autorités afin de les anticiper et de mieux gérer leur conséquences.

Deux problèmes émergent au moment d'étudier ces deux phénomènes : (1) l'intégration des caractéristiques spatiales et temporelles des données dans le processus d'ECD ; et (2) le passage à l'échelle des méthodes proposées.

Si ces techniques ont été adaptées aux deux cas d'études précédemment décrits, elles sont généralisables à d'autres phénomènes spatio-temporels.

1.5 Principales contributions

Ces dernières années, la communauté scientifique dans le domaine de l'extraction de connaissances, s'est intéressée à l'extraction de motifs spatio-temporels. Ce type de motifs capture les caractéristiques spatiale et temporelle des données en les généralisant sur différentes structures (e.g. co-locations, sous-graphes, etc.). Comme nous allons le voir dans le chapitre suivant, de nombreux travaux s'intéressent aux interactions spatio-temporelles des données. Néanmoins, les relations spatiales existantes entre les entités étudiées et la répercussion de sa prise en compte sur l'analyse des données reste mal connue.

Dans ce manuscrit de thèse, nous nous focalisons sur l'impact des composantes spatiales et temporelles des données sur le processus d'ECD. Nous avons étudié deux approches sémantiquement différentes d'extraction de motifs incluant des caractéristiques spatiales associées aux données spatio-temporelles :

1. **Les motifs spatialement fréquents :** nous nous focalisons sur le pré-traitement des données afin d'intégrer la composante spatiale décrite pour les entités spatiales. Pour cela nous utilisons différentes approches de spatialisation permettant de prendre en compte les relations spatiales existantes entre ces entités (e.g. des stations d'échantillonnage localisées le long d'une rivière).

6. "L'eau, toujours source de vie ? L'état réel des eaux et des données sur l'eau", *Anne Spiteri. (2011)*. Rapport téléchargeable sur http://eau-evolution.fr/

2. **Les motifs spatio-séquentiels :** nous proposons un nouveau type de motifs appelé *motifs spatio-séquentiels*. Ce type de motifs représente l'évolution temporelle d'un ensemble de caractéristiques décrivant une zone (e.g. ville, rivière, etc.) en prenant en compte leur voisinage proche.
3. **Deux algorithmes d'extraction de motifs spatio-séquentiels :** pour extraire les motifs spatio-séquentiels, nous avons défini, implementé et expérimenté deux algorithmes génériques basés sur deux stratégies souvent utilisées dans le domaine de fouille de données : parcours en profondeur et parcours par niveau.
4. **Deux mesures d'élagage adaptées à nos algorithmes et à notre problématique :** nous avons proposé deux mesures d'élagage anti-monotones permettant de filtrer les motifs dits "non fréquents".
5. **Deux mesures de qualité :** nous avons défini deux mesures de qualité afin d'évaluer la pertinence des motifs extraits. Ces deux mesures sont adaptées à nos deux types de motifs et permettent de faire ressortir les motifs "intéressants", tout en filtrant ceux fortement contredits par les données d'origine.
6. **Un prototype de visualisation :** nous avons développé un prototype de visualisation permettant d'afficher les deux types de motifs proposés dans cette thèse. Cette dernière étape permettra à l'expert une interprétation plus facile des résultats.

Ces différentes propositions ont été validées par des expérimentations sur des jeux de données synthétiques et réels. Deux jeux de données réelles ont été utilisés dans cette thèse. Le premier issu de la DASS de Nouvelle Calédonie, l'Institut Louis Pasteur, l'IRD et l'Université de Nouvelle Calédonie (Convention 2010) pour l'étude des épidémies de dengue. Le deuxième issu du projet *Fresqueau* (ANRII_MONU14), mené par l'ENGEES à Strasbourg et TETIS à la Maison de la Télédétection de Montpellier pour l'étude de la qualité de l'eau des rivières.

Ce mémoire s'inscrit dans le cadre d'une thèse cofinancée par la Région Languedoc Roussillon et l'Université de Nouvelle Calédonie.

1.6 Organisation du manuscrit

Ce manuscrit s'organise en six chapitres. Le Chapitre 2 présente un panorama des travaux étudiant cette problématique d'analyse des données spatio-temporelles. Nous avons identifié trois catégories : les trajectoires, les motifs spatiaux et les motifs spatio-temporels. Dans le Chapitre 3, nous présentons deux approches d'extraction de motifs spatio-temporels. La première capture les caractéristiques spatiales des données avant d'extraire des motifs tandis que la deuxième présente un nouveau type de motifs spatio-temporels appelé *motifs spatio-séquentiels* (S2P) qui permet l'étude de l'évolution d'un ensemble d'évènements décrivant une zone et leur entourage. Le Chapitre 4 s'intéresse à la découverte de motifs dits "intéressants" sélectionnés grâce à deux mesures de qualité.

Ces deux mesures, permettent de filtrer les motifs les moins contradictoires par rapport aux données. Les expérimentations sur des données réelles et synthétiques ont été réalisées pour tester la pertinence et l'efficacité de nos propositions. Ces expérimentations sont présentées dans le Chapitre 5. Dans le Chapitre 6 nous proposons un outil de restitution et de visualisation de motifs appelé *S2PViewer*. Enfin, nous concluons et proposons les différentes perspectives de recherche associées.

Chapitre 2

Etat de l'art

Préambule

Dans ce chapitre nous allons présenter un état de l'art détaillé des travaux étudiant l'extraction de motifs dans des bases de données spatio-temporelles. Nous montrerons les spécificités de ces bases et les limites des approches actuelles dédiées au suivi des trajectoires d'objets mobiles ou à l'études d'évènement localisés, vis à vis des objectifs que nous nous sommes fixés dans le Chapitre 1.

2.1 Introduction

L'explosion du nombre de sources d'informations spatiales et de systèmes d'information géographique (GIS), a fait émerger de nouveaux défis en matière d'analyse de données. En effet, ces informations représentent souvent des phénomènes complexes, qui sont difficiles à appréhender par la seule lecture des données. La mise en place de moyens analytiques permettant à ces données de faire sens pour les experts est un enjeu critique.

Yuan [2009] aborde la découverte de connaissances dans des bases de données géographiques et vise plus précisément à modéliser la dynamique spatio-temporelle des phénomènes sous-jacents. Le terme dynamique caractérise ici *le travail des forces qui entraînent le comportement d'un système et de leurs composants, individuelle et collective.* Par exemple, la dynamique d'un phénomène météorologique comme une tempête correspond aux interactions entre la température, la pression atmosphérique et tout autres facteurs contribuant à son évolution. La modification d'un des composants du système peut déclencher des variations dans les interactions entre les composants et finalement,

des variations dans la dynamique globale du système. Ces composants peuvent être très variés tels que des conditions climatiques (humidité, chaleur, précipitation, etc.) ou liés à la présence d'accidents géographiques (e.g. des montagnes, des rivières, etc.).

Face à ces questions, les méthodes de fouille de données spatio-temporelles visent à apporter des solutions pertinentes via l'identification, sans hypothèse *a priori*, de relations entre variables et évènements, caractérisés dans l'espace et dans le temps. Toutefois, nous allons montrer dans cet état de l'art que les méthodes de fouille de données spatio-temporelles actuelles sont limitées face à ce type de problématique.

Ce chapitre est organisé de la façon suivante. Tout d'abord, nous présentons les deux principaux types de bases de données spatio-temporelles : celles associées aux trajectoires d'objets mobiles et celles associées à l'étude d'évènements localisés. Ensuite, nous présentons les travaux qui ont été menés dans le cadre de l'extraction d'information dans ces deux types de bases de données spatio-temporelles.

2.2 Base de données spatio-temporelles

Comme expliqué dans la Section 1.3, une base de données spatio-temporelles contient des informations que l'on peut caractériser selon une dimension spatiale et une dimension temporelle. Plus formellement, on peut définir ces bases de données de la manière suivante :

Définition 2.1 *Une base de données spatio-temporelles est un ensemble structuré d'informations défini comme un triplet* $BD = \{D_S, D_T, D_A\}$ *où* D_T *est la dimension temporelle,* D_S *la dimension spatiale et* $D_A = \{D_{A_1}, D_{A_2}, \ldots D_{A_p}\}$ *l'ensemble des dimensions qui décrivent les autres attributs.*

La *dimension temporelle* est associée à un domaine de valeurs ordonnées dénoté $dom(D_T) = \{T_1, T_2, \ldots, T_t\}$ où T_i pour $i \in [1..t]$ est une *estampille temporelle* et $T_1 < T_2 < \ldots < T_t$. La *dimension spatiale* est associée à un domaine de valeurs dénoté $dom(D_S) = \{Z_1, Z_2, \ldots Z_l\}$ où chaque Z_i pour $i \in [1..l]$ est une *instance* matérialisant une zone, un objet ou un évènement localisé. Chaque *dimension d'analyse* D_{A_i} pour $i \in [1..p]$ est associée à un domaine de valeurs dénoté $dom(D_{A_i})$. Dans ces domaines, les valeurs peuvent être ordonnées ou non.

Deux types de bases de données spatio-temporelles sont principalement considérées : celles étudiant les trajectoires d'objets qui évoluent dans l'espace et le temps (e.g. des trajectoires d'oiseaux, d'avions) et celles étudiant les dynamiques spatiales et temporelles d'évènements (e.g. évolution de l'érosion dans une région ou propagation d'une épidémie dans une ville). Nous allons décrire les données associées à ces deux types de bases et donner des exemples d'interrogations qu'il est possible de formuler à partir de telles données. Pour cela, nous allons illustrer les concepts sur deux exemples : les déplacements d'orages et les phénomènes climatiques localisés dans une région.

2.2.1 Trajectoires d'objets mobiles

Les trajectoires peuvent être vues comme des ensembles de points localisés dans l'espace et le temps (*time-stamped coordinates*). $T = \langle (t_1, x_1, y_1), ..., (t_n, x_n, y_n) \rangle$ est une trajectoire, i.e., un ensemble de positions (x_i, y_i) de l'objet étudié aux temps t_i. La Figure 2.1 illustre la trajectoire d'un objet (un orage) à trois dates successives.

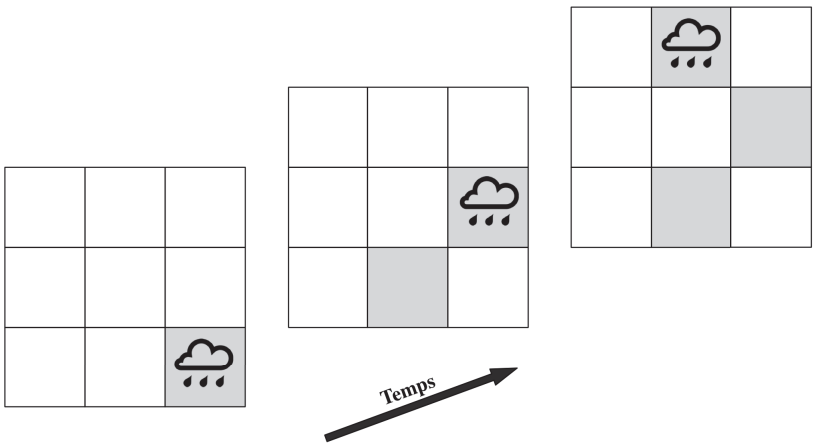


FIGURE 2.1 – Exemple de trajectoire de tempêtes

La dimension temporelle D_T des trajectoires dépend de la manière dont on capture les positions des objets : (1) les positions des objets sont enregistrées à intervalles de temps réguliers, e.g. toutes les 5 secondes le GPS d'un véhicule enregistre la position de ce véhicule (*time-based recording*) ; (2) un enregistrement de la position se fait à chaque fois que l'objet change de position, e.g. à chaque fois que l'utilisateur d'un smartphone se connecte sur facebook dans un lieu "connu" (*change-based recording*) ; (3) un enregistrement se fait quand l'objet se rapproche d'une position, e.g. à chaque passage près d'un capteur (*location-based recording*) ; et (4) l'enregistrement se fait pendant des évènements prédéfinis, e.g. à chaque appel depuis un téléphone mobile (*event-based recording*). Des caractéristiques comme la durée ou la périodicité sont généralement associées à la composante temporelle.

La dimension spatiale D_S des trajectoires dépend de l'information que l'on garde pour localiser l'objet étudié. Celui-ci peut être positionné sur une carte à l'aide de coordonnées géographiques (e.g. latitude, longitude), polaires ou toutes autres informations (même textuelle). Des caractéristiques comme la direction ou le type de déplacement (en ligne droite, curviligne, circulaire, etc.), les points d'inflexions, peuvent également être associés à la composante spatiale.

L'étude des trajectoires d'objets mobiles est initialement centrée sur le déplacement des objets dans l'espace (autrement dit, sur les dimensions spatiale et temporelle). Par la

suite, les dimensions d'analyse D_A ont été enrichies avec des informations statiques sur les objets mobiles eux mêmes (e.g. le diamètre de la tempête, les rafales, etc.).

Le Tableau 2.1 montre un exemple de base de données d'objets mobiles. L'attribut *Date* correspond à la composante temporelle et les attributs x_i, y_i à la composante spatiale.

TABLE 2.1 – Exemple de base de données d'objets mobiles

Objet	**Date**	x_i	y_i
$Objet_1$	t_1	x_1	y_1
$Objet_1$	t_2	x_2	y_2
$Objet_1$	t_3	x_3	y_3
$Objet_2$	t_1	x_4	y_4
$Objet_3$	t_1	x_5	y_5
...	...	...	...

À partir d'une telle base, un exemple typique de question (requête) associée à de l'information spatiale serait : *Au cours de l'année passée, combien de tempêtes se sont déplacées de la ville de Montpellier vers la ville de Nîmes ?* On peut utiliser ces informations pour déclencher une alerte *à chaque fois que des rafales dépassent les* 50 *km/h* ou établir une requête prédictive : *combien de tempêtes vont atteindre la ville de Montpellier dans les* 2 *jours à venir ?*

2.2.2 Evènements localisés

Un deuxième type de base de données spatio-temporelles consiste à stocker des informations sur des évènements (ou des objets) localisés dans l'espace et le temps. On considère ici des zones localisées dans l'espace, dans lesquelles se déroulent des évènements dont on connait l'estampille temporelle. $T = \langle (z_i, t_i, e_i) \rangle$ décrit un évènement e_i se déroulant dans une zone z_i au temps t_i. La Figure 2.2 décrit une base de données météorologiques répertoriant des évènements météo (e.g. pluie, rafales, basse température) associés à 9 zones sur trois temps consécutifs.

De même que précédemment, la composante temporelle de ces évènements localisés dépend de la manière dont on capture l'apparition de ces évènements : (1) tous les évènements sont enregistrés à intervalle de temps réguliers dans une zone, e.g. toutes les 5 minutes (*time-based recording*) ; (2) un enregistrement se fait à chaque fois qu'une dimension d'analyse particulière dépasse un seuil, e.g. à chaque fois que la température dépasse un seuil (*alert-based recording*) ; et (3) l'enregistrement se fait par identification du matériel à des dates prédéfinies, e.g. l'unité P0221 fait un enregistrement le 04/02/2013 entre 13h00 et 14h10 (*identifier-based recording*).

La dimension spatiale D_S correspond à la définition des zones dans lesquelles se déroulent les évènements. En fonction de la granularité spatiale, la zone z_i peut correspondre

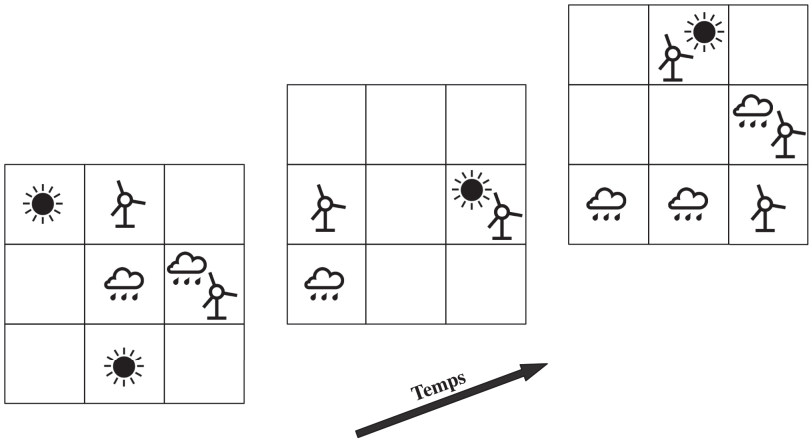


FIGURE 2.2 – Exemple d'évènements localisés

à une région (e.g. une région administrative), à un objet (e.g. une montagne) ou à un point si l'évènement est très localisé (e.g. une localisation via GPS). Ces zones peuvent être réparties dans l'espace de manière plus ou moins homogènes. Par exemple, elles peuvent être représentées sous la forme d'une grille aux formes variées (pavage carré, octogonal, etc.) ou sous la forme de polygones ayant des frontières communes (e.g. les quartiers d'une ville). Dans d'autres cas, les zones sont réparties de manières éparses dans l'espace, avec des possibilités de chevauchement et ont des formes très variées (polygones, lignes, points). On peut considérer par exemple un découpage administratif (e.g. une région), qui se superpose à un découpage géographique (e.g. une montagne, une plaine). Ces zones peuvent être étudiées en fonction de différentes proximités spatiales (e.g. à côté de, près de, en amont de, etc.).

Les dimensions d'analyse D_A correspondent aux différents types d'évènements et d'objets étudiés. Chaque dimension est associée ensuite à un domaine de valeurs représentant les différentes propriétés/caractéristiques des évènements et d'objets étudiés (e.g. pluie faible/forte).

Le Tableau 2.2 illustre un exemple de base répertoriant des évènements météorologiques. Cette base de données associe des évènements météo à trois villes sur deux jours consécutifs. Le tableau contient la température, les précipitations et la vitesse des rafales en *km/h*. Nous avons donc $D_T = \{Date\}$, $D_S = \{Ville\}$ et $D_A = \{Température, Précipitation, Rafales\}$. Le domaine de la dimension temporelle est $dom(D_T) = \{22/07/13, 23/07/13\}$ avec $22/07/13 < 23/07/13$. Le domaine de la dimension spatiale est $dom(D_S) = \{Z_1, Z_2, Z_3\}$. Finalement, le domaine de la dimension d'analyse *Température* est $dom(Température) = \{T_l, T_m, T_s\}$ et de la dimension d'analyse *Rafales* est $dom(Rafales) = \{55, 75\}$. Il est possible également d'associer des informations statiques aux zones (e.g. une surface, un relief, etc.).

TABLE 2.2 – Exemple de base de données d'évènements localisés

Ville	Date	Température	Précipitation	Rafales
Z_1	22/07/13	T_m	P_m	-
Z_1	23/07/13	T_m	P_m	55
Z_2	22/07/13	T_m	P_m	-
Z_2	23/07/13	T_b	P_m	75
Z_3	22/07/13	T_b	P_m	55
Z_3	23/07/13	T_m	P_h	-
...	...	...	...	...

À partir d'une telle base, il est possible de formuler des requêtes telles que : *Quelle zone a été la plus touchée pour une tempête les deux derniers mois ?* Des requêtes prédictives peuvent être posées telles que : *Dans quelle zone peut-on prévoir des fortes pluies ?* ou *S'il pleut au nord de Montpellier à midi, quelles seront les conséquences dans l'après midi au sud ?*

Dans la suite de ce chapitre, nous nous intéressons à ces deux types de bases des données et aux différentes méthodes permettant d'en extraire de l'information. Nous commençons par les méthodes dédiées au suivi des trajectoires puis nous continuons avec des méthodes dédiées aux bases d'évènements localisés.

2.3 Suivi de trajectoires d'objets mobiles

L'émergence des nouvelles technologies mobiles a entrainé la collecte de grandes quantités de données spatio-temporelles, permettant ainsi d'entrevoir de nouvelles applications notamment concernant le suivi de trajectoires. Par exemple, le projet GeoPKDD [Giannotti et Pedreschi, 2008] a étudié l'aménagement du plan de circulation de grandes agglomérations en fonction des déplacements des véhicules. L'analyse d'objets en mouvement a également comme domaines d'applications la géographie socio-économique, le sport (e.g. le déplacement des supporters de football au stade [Giannotti *et al.*, 2011]), l'analyse et le contrôle de la pêche, les prévisions météorologiques et l'analyse du mouvement (suivi de migration d'aigles [Li *et al.*, 2011]). La Figure 2.1 présente schématiquement une base de données de trajectoires. Le nombre de trajectoires étant important, l'un des objectifs des méthodes d'extraction est de trouver les séquences les plus pertinentes (généralement les plus fréquentes). Face à cette problématique, plusieurs approches ont été proposées dans la littérature. Nous nous focalisons dans la suite de cette section aux méthodes de fouille de données.

Dans [Mamoulis *et al.*, 2004; Cao *et al.*, 2005, 2007], les auteurs se sont intéressés à l'extraction de motifs périodiques dans de telles bases. Les objets étudiés (e.g. des orages) ont

la particularité de suivre approximativement la même route à intervalles de temps réguliers, e.g. *très fréquemment il y a des pluies saisonnières au début de l'été*. Cette approche résume l'ensemble des trajectoires d'un même objet par une seule séquence composée de segments. Les segments de trajectoires similaires sont regroupés en utilisant une fonction de similarité qui tient compte de la proximité spatiale, basée sur l'angle et la longueur spatiale des segments. Cette approche donne ainsi une meilleure abstraction des trajectoires et diminue la taille des données pour l'analyse. De plus, les auteurs relaxent la contrainte spatiale afin d'extraire un motif même s'il ne se répète pas exactement au même endroit. Pour extraire les trajectoires les plus fréquentes, ils ont utilisé un algorithme par niveaux dérivé d'*Apriori* et l'ont optimisé grâce à l'utilisation d'une nouvelle structure de données (substring tree). Ce travail a été validé sur une base de données de trajectoires de bus, où chaque séquence correspond aux déplacements d'un bus dans une journée.

Dans [Fisher *et al.*, 2005], les motifs étudiés sont des groupes d'objets partageant le même type de déplacement (direction, vitesse) à une date donnée dans une certaine région de l'espace. Cinq types de trajectoires basés sur le mouvement, la direction et la localisation sont proposés (convergence, rencontre, troupeau, leadership et récurrence). Les travaux présentés dans [Gudmundsson *et al.*, 2004] permettent de détecter les 4 premiers types de motifs définis dans [Fisher *et al.*, 2005] en utilisant des algorithmes de calcul approximatif. Les motifs spatio-temporels identifiés sont des sous-groupes d'objets ponctuels mobiles associés à un nombre suffisant d'éléments localisés dans une région et présentant un mouvement similaire (du point de vue de la direction, du but visé et/ou de la proximité). Un exemple de motifs extraits par ce type d'approche est : *un grand nombre de nuages annonçant de la pluie se sont déplacés ensemble vers le nord-est de Montpellier ce matin*.

Nanni et Pedreschi [2006] présentent une adaptation d'un algorithme de "clustering" basé sur la densité pour analyser les trajectoires d'objets en mouvement. Ils s'appuient sur la notion de distance entre trajectoires et la notion de densité. Dans le même article, les auteurs généralisent le "clustering" de trajectoires en se focalisant sur l'aspect temporel des données. Il s'agit de rechercher des intervalles de temps les plus significatifs, ce qui permet d'isoler les clusters (basé sur la densité) les plus intéressants. Par exemple, nous pouvons extraire l'information suivante : *Montpellier a subi de fortes précipitations le 22 mai, très probablement causées par des rafales provenant du sud et de basses pressions atmosphériques*.

Giannotti *et al.* [2007] proposent une extension du paradigme d'extraction de motifs séquentiels à l'analyse des trajectoires. Ils définissent les trajectoires comme des évènements fréquents depuis deux points de vue : en termes d'espace (les zones visitées lors des déplacements) et en termes de temps (la durée des déplacements). Ce travail est d'avantage axé sur des concepts de niveau supérieur au lieu de découvrir un motif impliquant un endroit spatial précis, une localisation générale est trouvée. Ces localisations générales sont appelées *régions d'intérêt* (*Regions-of-Interest* ou *RoI*). Les motifs fréquents de déplacement entre ces régions sont découverts par la suite. Par exemple, nous pouvons extraire des mo-

tifs du type : *des nuages annonçant de la pluie se sont déplacés de Montpellier vers Nîmes et au moins 70% ont touché la ville de Lunel*. Cette dernière ville est une région d'intérêt.

Plus récemment, Hai *et al.* [2012, 2013] ont proposé un "framework" basé sur trois étapes. Tout d'abord, ils ont développé une approche unificatrice pour extraire et gérer de multiples types de motifs représentant des trajectoires (e.g. des convois, des essaims, etc.). Ensuite, ils ont proposé un nouveau type de trajectoires appelées *rGpatterns* et ils ont proposé un algorithme efficace pour les extraire. Pour sélectionner les trajectoires, les plus intéressantes, les auteurs ont conçu une étape de compression basée sur le principe de *la longueur de description minimale* qui a été largement utilisé pour la sélection de modèles statistiques et dans les techniques de *machine learning*. Les auteurs ont fourni également un système de visualisation de trajectoires nommé *Multi_Move* qui est conçu pour extraire efficacement et automatiquement les différentes trajectoires et les visualiser à l'aide de *Google Earth*.

Nous allons maintenant détailler dans les deux sections suivantes, les travaux de la littérature relatifs à l'étude d'évènements localisés tels que définis dans la Section 2.2.2, que nous illustrerons sur l'exemple des phénomènes climatiques. Nous nous intéresserons à deux familles de motifs : les motifs spatiaux et les motifs spatio-temporels.

2.4 Motifs spatiaux

L'extraction de règles d'association spatiales et de motifs spatiaux a été largement étudiée ces dernières années dans des données géographiques et des SIG. Il y a deux familles d'approches : les approches multi-relationnelles et les approches basées sur les co-localisations (*colocations* en anglais) [Shekhar et Huang, 2001].

Lorsque les données sont composées de plusieurs tables relationnelles décrivant les objets et leurs relations spatiales, les techniques de fouille de données multi-relationnelles peuvent être utilisées pour extraire des modèles tels que des règles d'association multi-niveaux [Lisi et Malerba, 2004]. Dans ces travaux, les auteurs utilisent la programmation logique inductive pour définir des prédicats de relations spatiales. L'idée de base est de transformer la base spatiale en une base de données déductive, en pré-calculant toutes les relations spatiales. Ces dernières sont ensuite traduites sous forme de prédicats. Une base de données transactionnelles est alors produite. Une transaction est un ensemble de valeurs (appelées aussi caractéristiques) associé à un objet d'étude géo-référencé. L'information spatiale n'est donc plus explicitement présente dans ces données, mais elle est encodée dans la transaction. D'autres travaux se sont attaqués à la fouille de données spatiales en proposant des alternatives de fouille de données multi-tables [Chelghoum et Zeitouni, 2004] pour optimiser les calculs de jointure par utilisation d'index de jointures [Zeitouni *et al.*, 2000].

Les travaux de Koperski et Han [1995] et Bogorny *et al.* [2006] s'appuient sur le même type d'approche. Koperski et Han [1995] proposent une extraction des règles d'association

dans des bases de données géographiques en se focalisant sur une caractéristique spatiale de référence (e.g. les villes en bord de mer). Leur méthode énumère les voisinages de la caractéristique spatiale étudiée afin de "matérialiser" un ensemble de transactions correspondant aux instances de celle-ci. Ils ont utilisé les mesures standards de fréquence et de confiance après avoir calculé les distances et les relations spatiales. Un algorithme d'extraction d'itemsets est ensuite appliqué sur ces transactions. Il a permis ainsi de trouver les co-localisations liées à la caractéristique spatiale de référence (e.g. les villes de bord de mer en Colombie-Britannique sont souvent proches des Etats-Unis). Dans le même ordre d'idée, Bogorny *et al.* [2006] utilisent un ensemble d'objets géographiques de référence (les objets cibles) et codent de la même manière les relations spatiales. Les auteurs ont également introduit des connaissances expertes (dépendances connues) pour filtrer des motifs et éliminer les règles d'association redondantes.

L'approche basée sur les co-localisations se focalise sur les objets et leurs relations spatiales. Elle a été proposée par Shekhar et Huang [2001] avant d'être étendue dans [Huang *et al.*, 2004; Shekhar, 2006; Celik *et al.*, 2007, 2008; Lin et Li, 2009; Wang *et al.*, 2009]. Leurs travaux s'inspirent des résultats de Koperski et Han [1995] tout en affinant la définition de co-localisation et généralisant la méthode. L'objectif de cette approche est de trouver tous les sous-ensembles de propriétés (ou types d'évènements) fréquemment associés à des objets spatiaux voisins. Le motif {pluie,basseTempérature,rafales} dans la Figure 2.3 est un exemple de co-localisation qui traduirait le fait que des rafales sont aperçues souvent avec des pluies dès qu'il y a des basses températures. Il représenterait le fait que les propriétés *pluie, basseTempérature* et *rafales* sont souvent corrélées spatialement. Contrairement à l'approche précédente, toutes les caractéristiques et relations de voisinage sont considérées et les données ne sont pas pré-traitées. Une mesure d'intérêt anti-monotone a été introduite, appelée l'*indice de participation*, pour filtrer les co-localisations les plus importantes. Un algorithme par niveaux extrait les solutions.

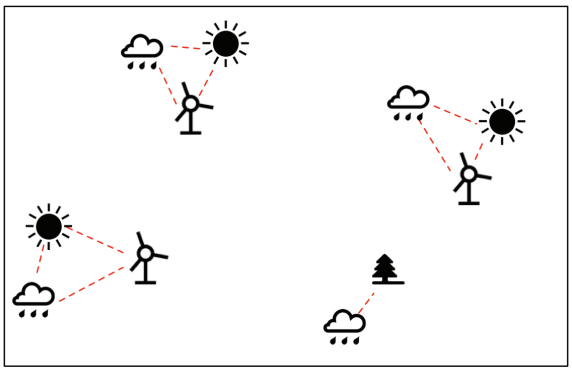


FIGURE 2.3 – Exemple de co-localisation

Dernièrement, Flouvat *et al.* [2010] et Selmaoui-Folcher *et al.* [2010, 2011] ont inté-

gré des connaissances du domaine directement dans le processus d'extraction de co-localisations.

Bogorny *et al.* [2006] se sont aussi intéressés à l'intégration des contraintes dans l'extraction de motifs spatiaux. Toutefois, leur travail s'appuie sur une approche orientée transactions qui ne prend en compte que partiellement les relations spatiales et fait un pré-traitement des données pour assimiler des contraintes expertes. L'idée de base est de considérer une caractéristique de référence, d'énumérer d'abord les voisinages pour construire un ensemble de transactions autour des instances de la caractéristique spatiale de référence. Les auteurs appliquent ensuite une méthode d'extraction d'itemsets, avec des contraintes expertes (e.g. éliminer tous ceux qui sont connus par l'expert), dans la base de données transactionnelles obtenue. Par exemple, la caractéristique de référence est *pluie*, une base de données transactionnelles générée considérant les caractéristiques de toutes les instances proches des instances de la référence, l'itemset fréquent *pluie, baseTempérature, vents* est extrait et considéré comme une co-localisation pertinente à la caractéristique de référence *pluie*.

Si ces motifs sont très intéressants en donnant la possibilité de révéler des associations entre des évènements et des lieux, ils ne permettent pas de matérialiser l'évolution et le déplacement de ces évènements. Par exemple, dans le cas de la météo, on ne peut capturer des informations sur des évolutions comme la transformation d'un nuage spécifique en orage près de courants d'air ascendants existant près des montagnes. Au contraire, l'objectif des motifs spatio-temporels, est d'étudier l'évolution et les interactions globales, dans l'espace et dans le temps, d'ensembles d'évènements.

2.5 Motifs spatio-temporels

L'objectif ici est différent de la fouille de trajectoires et de motifs spatiaux. En effet, il ne s'agit pas de suivre le déplacement d'objets en mouvement ou d'étudier des corrélations uniquement spatiales entre évènements, mais d'analyser globalement l'évolution et les interactions dans l'espace et dans le temps, d'ensembles d'évènements. Un exemple d'application est celui de l'*évolution spatio-temporelle d'un orage dans des différents régions grâce aux violents courants d'air verticaux entraînent l'humidité, les fragments de glace, les grêlons et les gouttelettes d'eau à l'intérieur du nuage*.

Dans cette catégorie de méthodes, Celik *et al.* [2006, 2008] ont généralisé le concept de co-localisations à des données spatio-temporelles. Les co-localisations spatio-temporelles représentent des ensembles de propriétés associées à des objets voisins dans l'espace et dans le temps. Plus précisément, ils représentent des instances spatialement proches pendant une fraction significative de temps. Une mesure d'intérêt monotone combinant prévalence spatiale et prévalence temporelle permet d'intégrer conjointement cette proximité spatiale et temporelle. Pour simplifier, seul les motifs apparaissant un grand nombre de fois dans un grand nombre de temps sont conservés. Dans

la Figure 2.4, les motifs {pluie,zoneUrbaine} et {pluie,basseTempérature,rafales} sont des co-localisations spatio-temporelles (les traits en pointillés représentent la relation de voisinage). Toutefois, {pluie,zoneUrbaine} aura une mesure d'intérêt plus forte que {pluie,basseTempérature,rafales} car il apparaît plus fréquemment. Pour extraire ces motifs, les auteurs utilisent une stratégie générer-tester-élaguer et un parcours par niveaux de type *Apriori*.

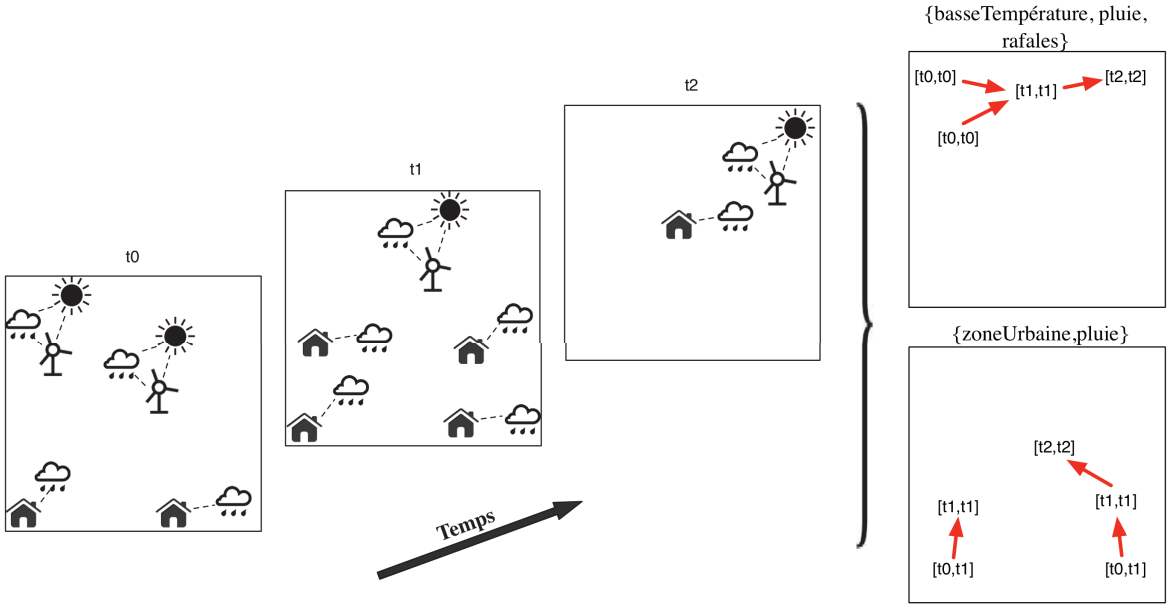


FIGURE 2.4 – Exemple de co-localisations spatiales *SPCOZ*

Le concept de co-localisations spatio-temporelles a aussi été étudié par Qian *et al.* [2009]. Les auteurs se sont intéressés à l'extraction des motifs représentant la propagation (la "trajectoire") de co-localisations spatiales appelées SPCOZ *(Spread patterns of spatio-temporal co-occurrences over zones)*. Ils suivent le déplacement d'éléments de propagation (*spread element*). Un élément de propagation est une co-localisation "fréquente" associée à une fenêtre temporelle. Dans la Figure 2.4, la co-localisation fréquente {pluie,zoneUrbaine} associée à l'intervalle $[t0,t1]$ est un élément de propagation. Les éléments de propagations combinés deux à deux constituent des arbres représentant la propagation d'un motif (SP-Tree ou Spread Pattern Tree). La Figure 2.4 illustre deux exemples de motifs SPCOZ : le SP-tree de {pluie,zoneUrbaine} et celui de {pluie,basseTempérature,rafales}. L'algorithme d'extraction recherche toutes les co-localisations fréquentes de taille 2 (par une méthode de type *Apriori*), les utilise pour construire les éléments de propagation pour tous les temps et génère les SP-Tree correspondants.

Dans [Yang *et al.*, 2005], les auteurs proposent un "framework" pour l'extraction de motifs spatiaux apparaissant fréquemment à différents temps et une extension permettant de visualiser certaines évolutions. Ils ont validé leur approche sur un jeu de données étudiant l'évolution de molécules et de vortex. Les motifs étudiés, appelés SOAP *(Spatial Object Association Pattern)*, sont représentés par des graphes. Chaque nœud correspond à une propriété et chaque arête représente une relation de voisinage. Ce travail partage des similitudes avec les co-localisations car cette approche permet aussi d'extraire des ensembles de propriétés associées à des objets voisins. Toutefois, à la différence de ces dernières, les auteurs considèrent des objets géométriques plutôt que des points pour les calculs de distances et de voisinage. De plus, ils permettent d'extraire trois autres types de configuration (cf., les exemples de la Figure 2.5) : étoile, séquence et *minLink*. Le dernier type permet de définir des SOAP plus généraux où seul le nombre minimum d'arêtes (*minLink*) associées à chaque nœud est fixé (c'est-à-dire le degré minimum des nœuds). Par exemple, si $\mathtt{minLink} = 1$, on obtient des SOAP de type étoiles, cliques ou séquences. Pour les séquences, les arêtes représentent une relation de voisinage et de direction. Par exemple, une arête (x, y) représente par exemple x *est voisin et au dessus de* y. Finalement, les auteurs montrent aussi comment utiliser (en post-traitement) les SOAP fréquents pour visualiser l'évolution d'un même ensemble de propriétés F. Pour cela, ils définissent la notion d'épisodes comme un ensemble d'instances associées à un intervalle de temps où le motif est apparu puis a disparu. Les épisodes de tous les SOAP fréquents associés à F sont recherchés (tous types confondus), puis ordonnés en fonction de leur date d'apparition. La séquence ainsi générée permet de visualiser l'évolution spatiale (étoile, clique, etc.) de l'ensemble des propriétés étudiées. Toutefois, ce post-traitement ne permet pas de prendre en compte les évolutions de forme des objets étudiés, ainsi que les éventuelles relations de cause à effet. En effet, seuls les changements globaux de "configuration" peuvent être observés. De plus, le nombre d'épisodes peut être important pour un même ensemble de propriétés, ce qui rend très difficile l'analyse de la séquence. L'intégration d'informations supplémentaires telle que la météorologie est également délicate.

Les séquences et plus généralement les graphes, ont été souvent utilisés et étendus au spatio-temporel pour représenter la propagation de phénomènes dans l'espace et dans le temps [Tsoukatos et Gunopulos, 2001; Wang *et al.*, 2004a, 2005; Huang *et al.*, 2008; Mabit *et al.*, 2011; Selmaoui-Folcher et Flouvat, 2011; Mohan *et al.*, 2012].

Tsoukatos et Gunopulos [2001] ont étendu les travaux sur les séquences d'itemsets (ensembles de caractéristiques environnementales) afin d'extraire des séquences représentant l'évolution dans le temps de zones d'études (e.g. des quartiers). La base de données considérée est constituée de séquences d'itemsets représentant l'évolution temporelle des différentes zones. Un algorithme effectuant un parcours en profondeur de l'espace de recherche est ensuite appliqué pour extraire les séquences les plus fréquentes (i.e., celles apparaissant dans le plus de zones). La Figure 2.6 illustre un exemple de séquences pouvant être extraites. Les auteurs ont également proposé une approche pour extraire les séquences fréquentes à une granularité spatiale plus élevée (e.g. région) en exploitant les séquences

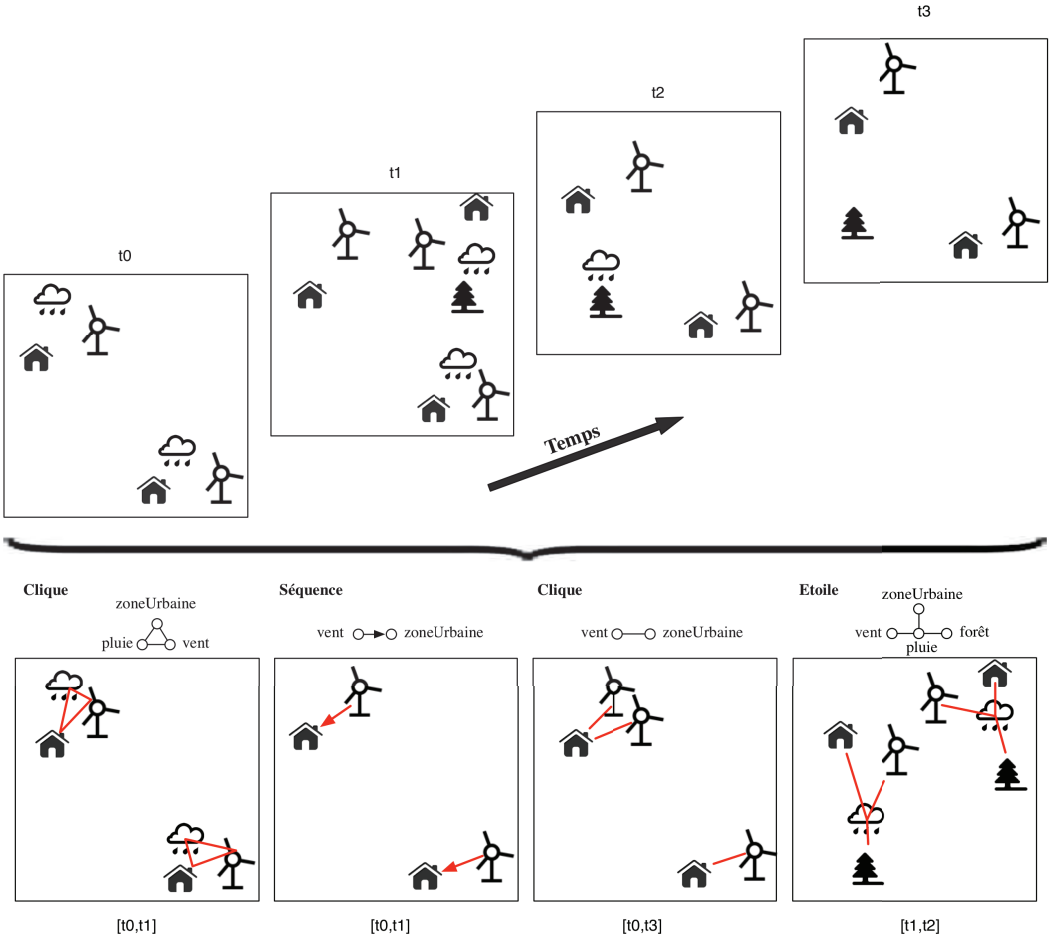
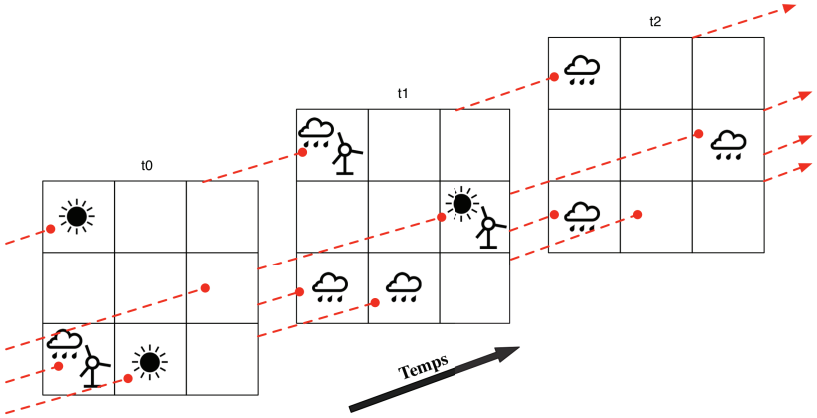


FIGURE 2.5 – Exemple de SOAP fréquents

fréquentes trouvées à une granularité plus faible (e.g. ville). Ils exploitent pour cela le fait que les séquences extraites à un niveau plus faible resteront fréquentes à un niveau de granularité plus élevé. La méthode recherche alors uniquement de nouvelles séquences fréquentes issues de l'agrégation spatiale.

Dans [Wang *et al.*, 2004a], les auteurs se focalisent sur l'extraction de séquences représentant la propagation spatio-temporelle d'évènements dans des fenêtres temporelles prédéfinies. Ils découpent la dimension temporelle en fenêtres d'une taille donnée (e.g. 4 jours), divisent l'espace sous la forme d'une grille et introduisent le concept de *flow pattern*. Un flow pattern est une séquence d'ensembles d'évènements de la forme $\langle E_1 \rightarrow \ldots \rightarrow E_k \rangle$ où E_i est un ensemble d'évènements de la forme *e(localisation)*, avec *e* un type d'évènements (e.g. pluie, vent). Chaque ensemble d'évènements est composé d'évènements spatialement voisins apparaissant au même temps. Deux ensembles d'évènements E_p et



<(températureElevée) (pluie)>, support = 3

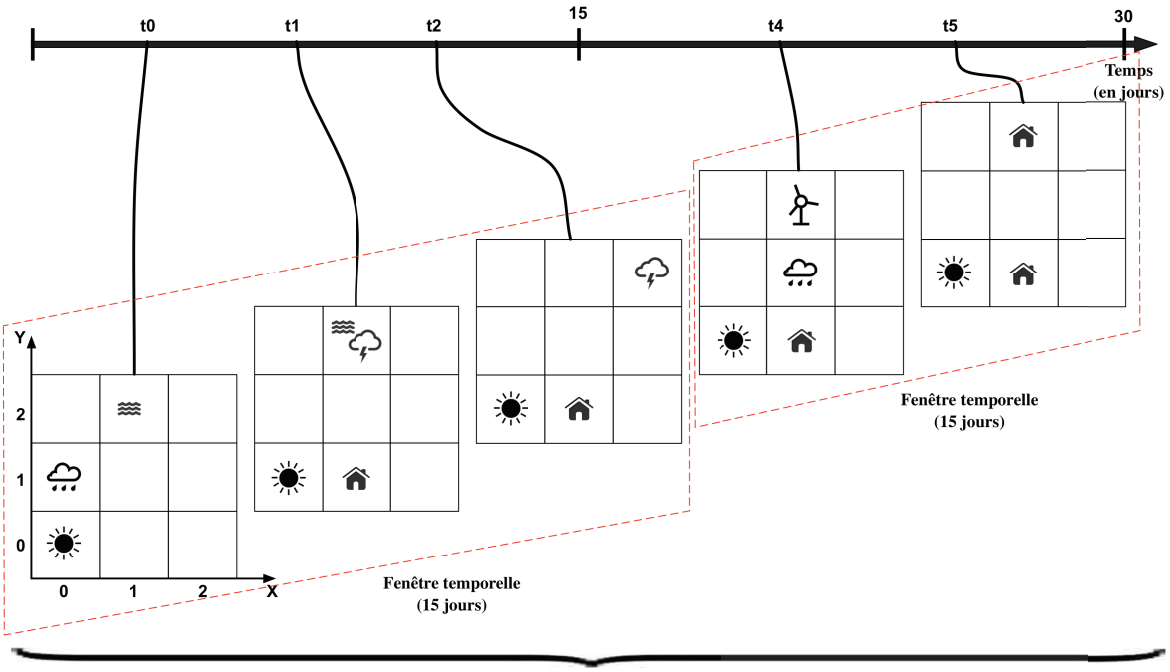
<(pluie, vent) (pluie)>, support = 2

<(pluie) (pluie)>, support = 2

<(pluie) (pluie) (pluie)>, support = 1

FIGURE 2.6 – Exemple de séquences représentant l'évolution des zones

E_q sont consécutifs dans la séquence, si leurs évènements appartiennent à la même fenêtre temporelle, s'ils sont tous voisins et qu'ils apparaissent à deux temps consécutifs. L'objectif de ce travail est de trouver les séquences d'évènements apparaissant un grand nombre de fois suivant les mêmes localisations. La Figure 2.7 montre un exemple de jeu de données et quelques flow patterns avec leur support (dans cet exemple, deux évènements sont voisins si leur distance euclidienne est inférieure ou égale à 1). Les évènements sont localisés par des coordonnées (X,Y) tel que l'évènement pluie(0,1) au temps t0. À titre d'exemple, le flow pattern ⟨{basseTempérature(0,0)}⟶{zoneUrbaine(1,0)}⟩ apparaît 3 fois. À l'opposé, le motif ⟨{basseTempérature(0,0),pluie(0,1)}⟶{dépôtEau(1,2),tempête(1,2)}⟩ a un support de 0 car les évènements {dépôtEau(1,2),tempête(1,2)} ne sont pas voisins de tous les évènements de {basseTempérature(0,0), pluie(0,1)} (dépôt d'eau et tempête du temps t1 ne sont pas voisins de basse température et pluie du temps t0). Il est aussi intéressant de noter que les motifs ⟨{basseTempérature(0,0), pluie(1,0)}⟶{zoneUrbaine(1,0)}⟩ et ⟨{basseTempérature(0,0),pluie(1,1)}⟶{zoneUrbaine(1,0)}⟩ sont considérés comme deux motifs différents bien que représentant des phénomènes similaires (seule la localisation de l'évènement pluie est légèrement différente). Autre point important, le support du motif ⟨{dépôtEau(1,2)}⟶{dépôtEau(1,2), tempête(1,2)}⟶{tempête(2,2)}⟶ {vent(1,2)}⟩ est égal à 0 car il n'est pas inclus dans une unique fenêtre temporelle. Pour extraire ces motifs, les auteurs appliquent une stratégie par niveau pour trouver les séquences de taille 1 et 2, puis utilisent les motifs fréquents trouvés comme point de départ à un parcours en profondeur de l'espace de recherche.



<{basseTempérature(0,0)} → {zoneUrbaine(1,0)}>, support = 3
<{basseTempérature(0,0),pluie(1,0)} → {zoneUrbaine(1,0)}>, support = 1
<{basseTempérature(0,0),pluie(1,1)} → {zoneUrbaine(1,0)}>, support = 1
<{dépôtEau(1,2)} → {dépôtEau(1,2),tempête(1,2)} → {tempête(2,2)}>, support = 1
<{dépôtEau(1,2)} → {dépôtEau(1,2),tempête(1,2)} → {tempête(2,2)} → {vent(1,2)} >, support = 0
<{vent(1,2)} → {zoneUrbaine(1,2)}>, support = 1

FIGURE 2.7 – Exemple de flow patterns

Dans un deuxième temps, Wang *et al.* [2005] étendent cette notion et définissent les motifs *spatio-temporels généralisés* (*Generalized spatio-temporal pattern*) comme des séquences de *relative eventsets*. Un *relative eventset* est un ensemble d'évènements dont la localisation est remplacée par un positionnement relatif à une localisation de référence. Un motif spatio-temporel généralisé est fréquent s'il a au moins *t-minsup* (support temporel) occurrences dans le temps et qu'il a au moins *s-minsup* (support spatial) occurrences dans l'espace (les localisations peuvent être différentes mais la localisation relative doit être identique). Pour extraire ces motifs, les auteurs proposent un nouvel algorithme appelé GenSTMiner qui utilise une approche dérivée de PrefixSpan [Mortazavi-Asl *et al.*, 2000]. Nous pouvons extraire, par exemple, la séquence traduisant le fait que *des températures élevées et de l'humidité apparaissent fréquemment dans une zone après l'apparition de forte pluie dans une zone voisine, pendant une semaine.*

Huang *et al.* [2008] se sont concentrés sur le problème d'extraction de séquences de propriétés représentant la propagation de certains types d'évènements. Ces séquences sont de la forme $\langle f_1 \rightarrow f_2 \rightarrow \ldots \rightarrow f_k \rangle$, où f_i est un type d'évènements. Cette approche permet donc d'étudier la propagation des évènements pris individuellement (sans prendre en compte leur environnement). Ce modèle considère deux évènements comme consécutifs s'ils sont spatialement proches (distance euclidienne inférieure à un seuil donné) et apparaissent dans la même fenêtre temporelle. Les auteurs ont également étudié d'autres relations de voisinage dépendant du temps. Ces relations permettent de représenter un rétrécissement de la zone d'influence d'un évènement (son voisinage) au cours du temps. Les auteurs proposent aussi une nouvelle mesure d'intérêt pour ces séquences car les mesures basées sur le support ne reflètent pas nécessairement un lien de cause à effet entre les évènements. Cette mesure n'étant pas anti-monotone, les auteurs proposent donc un nouvel algorithme Slicing-STS-Miner, pour extraire ces séquences, basé sur un traitement incrémental des différentes fenêtres temporelles et une extension des séquences à chaque étape. A défaut de l'anti-monotonie, ils exploitent une autre propriété du *sequence index* : si une séquence est intéressante, toutes les sous-séquences ayant le même préfixe sont intéressantes.

Dans la recherche de motifs spatio-temporels, nous trouvons aussi le travail de Mohan *et al.* [2012] qui ont proposé un nouveau motif appelé *motif spatio-temporel en cascade* (CSTP). Ce motif représente un sous-ensemble partiellement ordonné - par rapport aux motifs séquentiels qui sont entièrement ordonnés - de *types d'évènements* dont les instances sont situées ensemble et se produisent en série. Par exemple, un CSTP partiellement ordonné peut être : *une basse température et la présence de pluies conduisent à la formation de rafales de vent* (cf., Figure 2.8). L'ordre partiel sur un ensemble de type d'évènements peut être limité suite à la définition d'une relation de voisinage R dans laquelle la distance dans l'espace, dans le temps ou les deux est limitée par un seuil. Les auteurs représentent cette relation de voisinage R par un graphe orienté acyclique où les nœuds sont des objets spatiaux et les arcs représentent des relations de voisinage. Pour extraire les CSTP les plus intéressants, les auteurs s'inspirent des travaux de Shekhar et Huang [2001] et définissent le *Cascade Participation Index* comme mesure d'intérêt. Cette mesure anti-monotone est définie comme la valeur minimum des probabilités d'apparition d'un CSTP connaissant l'apparition de l'instance d'un type d'évènement qui participe au CSTP. Pour extraire des CSTP, les auteurs proposent un algorithme appelé CSTP Miner basé sur une stratégie du type *Apriori*. L'approche de Mohan *et al.* [2012] a été testé sur une base de données de crimes dans une grande ville pour extraire des relations existant entre des boites nocturnes, des agressions et des accidents de la route causés par l'alcool.

Le Tableau 2.3 résume les références bibliographiques organisées par type de motif extrait.

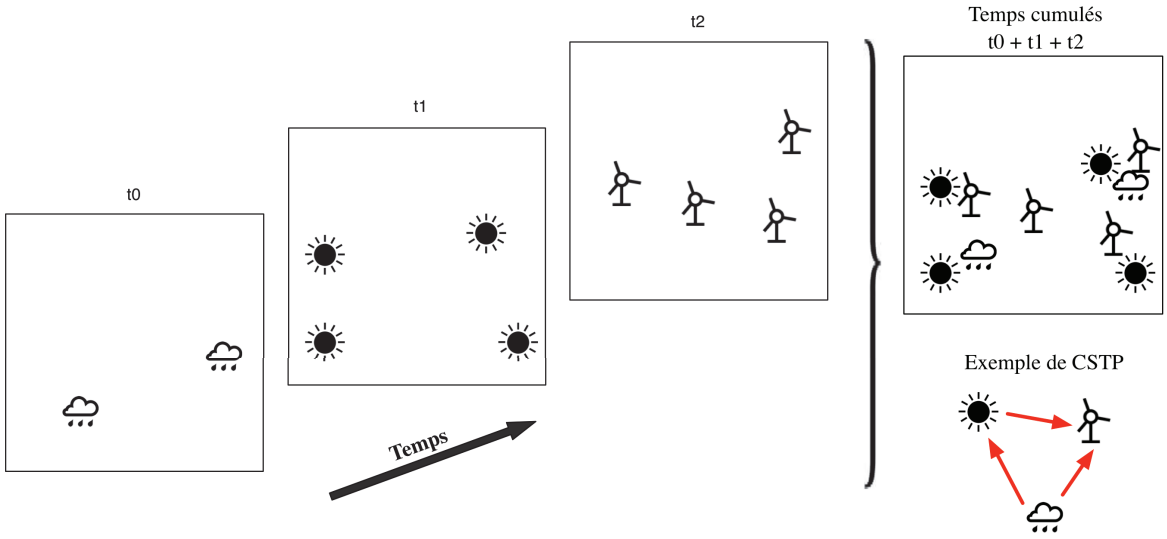


FIGURE 2.8 – Exemple de motif spatio-temporel en cascade (CSTP)

2.6 Discussion

Dans cet état de l'art, nous avons pu découvrir le nombre important de travaux étudiant l'extraction de motifs spatio-temporels qui sont résumés dans le Tableau 2.3 selon les trois approches adoptées dans cet état de l'art : suivi de trajectoires d'objets mobiles, motifs spatiaux et motifs spatio-temporels. Cependant, aucun de ces travaux ne permet d'étudier l'évolution d'un ensemble d'évènements décrivant une zone, tout en prenant en compte les caractéristiques (e.g. la position) et les évènements (e.g. basse température) des zones voisines.

Dans cette thèse, nous nous focalisons sur deux méthodes génériques d'extraction de motifs spatio-temporels. Pour cela, nous allons tout d'abord nous inspirer de la méthode proposée par Tsoukatos et Gunopulos [2001] dédiée à l'extraction de motifs séquentiels. Ces motifs étudient l'évolution au cours du temps d'un ensemble de caractéristiques. Cette approche semble donc particulièrement pertinente par rapport à notre problématique. Cependant, ces motifs ne représentent qu'indirectement la dimension spatiale des données. En effet, dans les travaux de Tsoukatos et Gunopulos [2001], les motifs ne permettent ni de représenter une évolution spatiale, ni une influence des évènements proches dans l'espace.

Nos contributions se focalisent sur les deux objectifs suivants : (1) intégrer la dimension spatiale dans la définition et l'extraction des motifs séquentiels. Différentes relations spatiales et topologiques seront étudiées (e.g. appartenance, distance, voisinage) ; et (2) proposer des méthodes qui passent à l'échelle.

TABLE 2.3 – Citations par type de motif extrait

Type de motif extrait	**Référence**
1. Trajectoires	[Giannotti et Pedreschi, 2008; Mamoulis *et al.*, 2004]
	[Cao *et al.*, 2005, 2007; Fisher *et al.*, 2005]
	[Gudmundsson *et al.*, 2004; Nanni et Pedreschi, 2006]
	[Giannotti *et al.*, 2007; Boulila *et al.*, 2010]
	[Hai *et al.*, 2012, 2013]
2. Motifs spatiaux	
2.1. Multi-relationnel	[Lisi et Malerba, 2004; Chelghoum et Zeitouni, 2004]
	[Zeitouni *et al.*, 2000; Koperski et Han, 1995]
	[Bogorny *et al.*, 2006]
2.2. Co-locations	[Shekhar et Huang, 2001; Huang *et al.*, 2004]
	[Shekhar, 2006; Celik *et al.*, 2007, 2008; Lin et Li, 2009]
	[Wang *et al.*, 2009; Koperski et Han, 1995; Bogorny *et al.*, 2006]
	[Flouvat *et al.*, 2010; Selmaoui-Folcher *et al.*, 2010, 2011]
3. Motifs spatio-temporels	
2.1. Co-locations spatio-temporelles	[Celik *et al.*, 2006, 2008; Qian *et al.*, 2009]
	[Yang *et al.*, 2005]
2.1. Séquences spatio-temporelles	[Tsoukatos et Gunopulos, 2001; Wang *et al.*, 2004a, 2005]
	[Huang *et al.*, 2008; Selmaoui-Folcher et Flouvat, 2011]
	[Mabit *et al.*, 2011; Cressie, 1993; Mohan *et al.*, 2012]

Nos deux propositions se situent à des étapes différentes du processus d'ECD : le prétraitement et la fouille comme nous les décrivons dans le chapitre suivant.

Chapitre 3

Fouille de motifs spatio-temporels

Préambule

Dans ce chapitre, nous présentons deux méthodes de fouille de données spatio-temporelles. La première permet l'extraction de "motifs spatialement fréquents". Ces motifs représentent l'évolution temporelle de zones homogènes préalablement regroupées selon différentes approches de spatialisation. La deuxième approche permet l'extraction d'un nouveau type de motifs appelé "motifs spatio-séquentiels". Cette deuxième approche permet d'étudier l'évolution au cours du temps d'un ensemble de caractéristiques représentant une zone et son entourage proche (i.e., zones voisines). Grâce à ces deux approches, nous extrayons deux types de motifs sémantiquement différents.

3.1 Motivations

Dans le contexte des bases de données multidimensionnelles, le mot *espace* est fréquemment utilisé. Le terme vient du latin *spatium* et correspond à une *certaine étendue superficielle dans laquelle une entité spatiale peut être placée*[1]. Il existe d'autres connotations plus spécifiques à certains domaines. Par exemple, dans les sciences du vivant, Johnston [2003] définit l'espace comme *les endroits sélectionnés par les humains pour y habiter, construire leurs maisons, leurs lieux de travail, etc.* Dans le domaine de la géographie, Clark et Clark [1993] définissent l'espace comme *un volume, une surface ou une longueur*

1. Oxford Dictionary of English 3rd. edition Copyright 2010 by Oxford University Press.

qui peut être occupée par quelque chose ou qui peut être vide. Nous pouvons citer comme exemple d'espace une étendue d'eau, une région administrative, la distance entre deux routes, etc. De manière générale, nous utiliserons le terme d'*entité spatiale* pour désigner des espaces.

On distingue généralement trois types de relations entre les entités spatiales : les relations topologiques, les relations de distance et les relations d'orientation.

– les **relations topologiques** décrivent les relations permettant de positionner globalement différentes entités spatiales entre elles (e.g. est connecté à, est adjacent à, est le début de, contient, etc.). La topologie assure la cohérence géométrique et la logique spatiale des entités, caractéristiques indispensables pour leur exploitation, en particulier pour des fonctions d'analyse spatiale. Lorsque nous observons le monde qui nous entoure, nous en percevons les éléments de manière globale. Par exemple, une chaise est *à côté* d'une table *dans* une pièce. Les entités sont vues dans leur contexte et la notion de voisinage est implicite. Les principales relations topologiques sont l'adjacence (contiguïté), la connectivité, l'inclusion et l'intersection. Par exemple, la parcelle de M. Dupont partage une frontière avec la parcelle de M. Rodriguez (adjacence). Le fleuve le Lez est relié à la Méditerranée (connectivité). Un département se situe dans une région (inclusion). Le confluant se situe à l'intersection du Tarn et de la Garonne (intersection). Les relations topologiques peuvent être représentées graphiquement. Par exemple, la Figure 3.1 montre l'entité spatiale A qui se trouve à côté de l'entité spatiale B ;

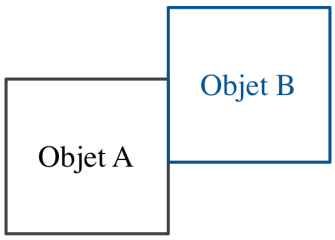


FIGURE 3.1 – Relation d'adjacence entre deux entités spatiales

– les **relations de distance** sont généralement basées sur la distance euclidienne entre deux entités spatiales. Soit *dist* une fonction de distance et p un prédicat de comparaison ($>, <, >=, <=$ ou $=$). Soit U un nombre réel et soient A et B deux entités spatiales, la relation de distance est notée par $dist(A,B)pU$ [Bogorny *et al.*, 2005]. Par exemple, Si A et B sont deux points, la fonction $dist(A,B) = 23,5$ peut être interprétée comme *il existe 23,5 unités de distance entre les points* A *et* B. Dans cet exemple, les points peuvent être définis à l'aide de leurs coordonnées cartésiennes (x,y). Dans la Figure 3.2, l'entité spatiale A se trouve à une certaine distance de l'en-

tité spatiale B. Cette distance a été mesurée depuis le centre de chaque entité spatiale par rapport au bord ;

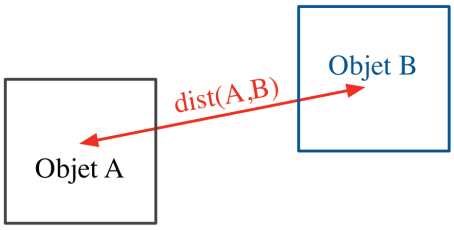


FIGURE 3.2 – Relation de distance entre deux entités

– les **relations d'orientation** expriment comment deux entités spatiales sont situées dans l'espace, l'une par rapport à l'autre ou par rapport à une entité spatiale de référence. Par exemple, la parcelle de M. Dupont se trouve au sud de celle de M. Rodriguez. Dans la Figure 3.3, l'entité spatiale A se trouve à l'ouest de l'entité spatiale B. On peut noter que les entités spatiales ne sont pas nécessairement collées.

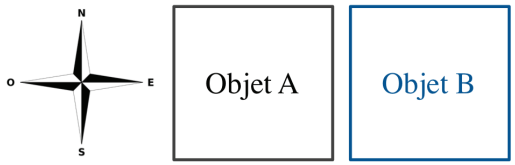


FIGURE 3.3 – Relation d'orientation entre deux entités

Ces trois relations sont liées à l'idée de groupement ou *géo-agrégation* qui intuitivement cherche à regrouper des entités spatiales proches selon ces relations. La géo-agrégation peut être définie comme *le processus de regroupement des entités spatiales appartenant à un même lieu, dans une aire de dimension restreinte ou qui suivent un ordre spécifique*[2]. Par exemple, les peuples se forment par agrégation de migrants, des quartiers par agrégation de maisons, etc. Dans le processus de géo-agrégation, nous supposons qu'il existe des similitudes entre les entités spatiales rapprochées. Par exemple des entités spatiales peuvent être regroupées en considérant : (1) une agrégation statistique qui regroupe des unités semblables par leurs attributs ou (2) une agrégation géographique qui permet de regrouper, par contiguïté, des unités géographiques en unités de niveau d'agrégation supérieur.

Toutes ces configurations spatiales et ses opérations (e.g. inclusion, appartenance, etc.) sont caractéristiques des données spatiales. La principale spécificité de la fouille de données spatiale est de les expliciter. À présent, la question est : **comment inclure la notion**

2. Encyclopédie électronique Hypergeo http ://www.hypergeo.eu/

de spatialisation dans le processus d'extraction de connaissances à partir des données ? Une première approche consiste à pré-traiter les données pour inclure des caractéristiques spatiales avant le processus de fouille de données. C'est ce que nous allons détailler dans la Section 3.2. Une deuxième approche consiste à définir de nouveaux types de motifs à rechercher dans les données qui expriment les relations spatiales. C'est ce que nous allons traiter dans la Section 3.3.

Pour illustrer les définitions dans cette section, nous allons utiliser la base de données spatio-temporelles *stDB* décrite dans le Tableau 1.1. Ce tableau représente l'évolution de quatre évènements météorologiques (la température, l'humidité, les précipitations et des rafales de vent) pour trois zones au cours de trois jours (cf., Figure 1.2).

Nous pouvons transformer les données du Tableau 1.1 en une base de séquences en regroupant les évènements météorologiques par date et par zone. Cette base de séquence représente donc l'évolution temporelle d'un ensemble d'évènements représentant une zone (voir Section 3.2.4).

Dans la section suivante, nous décrirons notre première approche.

3.2 Motifs spatialement fréquents

Une base de données spatio-temporelles est composée d'une dimension spatiale et d'une dimension temporelle. Comme discuté dans l'état de l'art du Chapitre 2, il est difficile d'étudier ces deux dimensions conjointement.

Une partie des approches considèrent que les phénomènes spatio-temporels peuvent être étudiés grâce à des séquences d'évènements évoluant au cours du temps. Ils recherchent généralement les séquences d'évènements fréquents [Agrawal et Srikant, 1995]. Ces séquences ne considèrent les changements que du point de vue "temporel". Il est donc intéressant d'inclure des caractéristiques spatiales associées aux entités (e.g. position, distance, etc.) dans le processus d'ECD de façon à extraire des séquences capturant ces caractéristiques.

Dans ce contexte, nous nous focalisons sur l'étape de pré-traitement du processus général d'ECD. Nous proposons de regrouper les entités spatiales stockées dans une base de données spatio-temporelles en utilisant des relations spatiales topologiques, de distance ou d'orientation. Ce pré-traitement permet d'étudier l'impact de la "spatialisation" des informations dans le processus de fouille de données. Une fois les données pré-traitées, nous utilisons un algorithme "classique" d'extraction de motifs séquentiels pour identifier les séquences fréquentes. Les motifs ainsi extraits sont appelés *motifs spatialement fréquents*. Dans cette section, nous décrivons les différentes étapes de ce processus.

3.2.1 Processus d'extraction de motifs spatialement fréquents

Notre approche se déroule en quatre étapes : (1) sélection des données ; (2) décomposition spatiale et agrégation ; et (3) fouille de données spatio-temporelles en utilisant un algorithme classique d'extraction de motifs séquentiels. Ce processus global est illustré dans la Figure 3.4.

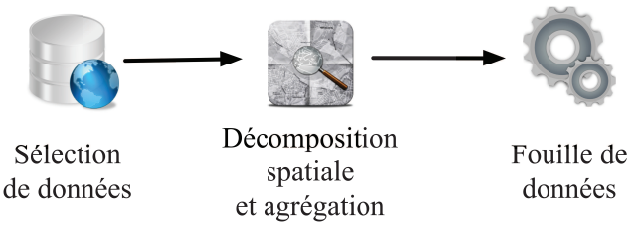


FIGURE 3.4 – Processus d'extraction de connaissances sur des données spatiales

Dans cette section, nous nous focalisons sur l'étape de pré-traitement qui est le cœur de cette approche. La décomposition spatiale et l'agrégation sont des étapes préalables au traitement des données spatiales. Les données sont regroupées dans différentes séquences spatiales selon les relations existantes entre les entités spatiales qui les supportent. Par exemple, certaines zones sont regroupées selon leur proximité et d'autres sont décomposées si elles ne partagent pas de caractéristiques communes. Grâce à cette transformation, les caractéristiques spatiales des données sont intégrées dans le processus d'ECD.

Les séquences spatiales résultant sont ensuite utilisées en entrée de l'étape de fouille de données, ce qui va nous permettre d'extraire des séquences spatialement fréquentes à l'aide d'un algorithme d'extraction de motifs séquentiels classique (e.g., [Mortazavi-Asl *et al.*, 2000]). Les motifs ainsi extraits représentent les évolutions temporelles spatialement fréquentes des zones.

La section suivante, détaille notre proposition pour intégrer la dimension spatiale afin de découvrir des séquences spatio-temporelles.

3.2.2 Pré-traitement des données

Les étapes de pré-traitement ont un fort impact sur les résultats du processus de fouille. La préparation des données est donc une étape primordiale souvent répétitive et les modèles découverts dépendent du type et de la qualité de l'ensemble des données en entrée. Il est important de noter que cette étape consomme de 60 à 80% du temps consacré à l'ensemble du processus d'ECD [Adriaans, 1996].

Il existe dans la littérature de nombreuses approches de pré-traitement des données. Plusieurs techniques dédiées aux données spatio-temporelles ont été étudiées dernièrement [Elias, 2003; Mennis et Liu, 2005; Qi et Zhu, 2003]. Chaque référence a ses propres

objectifs, comme la classification spatiale, les règles d'association spatiales ou la découverte de connaissances.

Dans [Bogorny *et al.*, 2005], le pré-traitement est utilisé pour intégrer des informations spatiales avant l'étape de fouille de données. Les données spatiales sont converties en prédicats spatiaux du type $P_1 \wedge P_2 \wedge \ldots \rightarrow Q_1 \wedge Q_2 \wedge \ldots$ où au moins un des P ou Q est un prédicat spatial (e.g. contienne(arbres), appartient(ville), croise(coursDeEau), etc.) et $\wedge$ est l'opérateur logique *and*. Grâce à cette transformation, un algorithme de fouille de données classique peut ensuite être exécuté pour extraire des structures spatiales.

La granularité est un autre critère important lorsque l'on pré-traite des données spatiales. Par exemple, si l'objectif est d'étudier les changements dans les données produites par des stations météorologiques, un moyen d'extraire des motifs spatiaux est d'agréger l'information spatiale pour plusieurs stations afin d'étudier des changements dans des zones homogènes (e.g. des stations localisées dans une ville). Dans le Tableau 1.1 (Chapitre 1), la zone Z_1 est représentée par la séquence d'évènements $T_b H_m P_m \rightarrow T_m H_m P_b \rightarrow T_b H_m P_m 55$ qui peut être interprétée comme l'apparition d'une température basse, d'humidité et de précipitations modérées, suivi d'une température et d'une humidité modérées ainsi que de faibles précipitations, pour finir avec une température basse, des précipitations et une humidité modérées, et des rafales de vent de 55 km/h. Dans [Koperski et Han, 1995; Tsoukatos et Gunopulos, 2001], les auteurs utilisent cette approche et ajoutent les caractéristiques spatiales à des ensembles (ou séquences) de valeurs.

À notre connaissance, aucune proposition n'a tenté d'extraire des motifs séquentiels à différents niveaux de granularité spatiale et de combiner ensuite leurs résultats pour obtenir des informations plus générales. Ces motifs doivent prendre en compte l'information spatiale associée aux entités étudiées et des hypothèses que l'on peut construire sur les phénomènes représentés par les données. Cette tâche n'est pas simple et nous oblige à remettre en question le processus d'ECD classique. Dans cette section, nous nous concentrons sur le comportement spatial du point de vue de la division de l'espace en utilisant différents niveaux de granularités spatiales construits grâce à l'opération de géo-agrégation. Cette tâche a été réalisée pour déduire des motifs plus généraux en moyennant des attributs des entités spatiales regroupées en zones homogènes.

3.2.3 Hypothèses de spatialisation

Schématiser un phénomène géographique complexe statique (n'évoluant pas dans le temps) ou au contraire dynamique, fait appel à de nombreuses échelles utilisées pour décrire des entités suivant une représentation qui n'est pas toujours évidente à construire. C'est justement à partir de cette modélisation que nous pouvons construire (par géo-agrégation) des entités spatiales plus complexes et de cette façon, extraire des motifs pour aboutir à une vision plus générale.

Concernant le groupement des entités, l'opération de géo-agrégation est toujours délicate. Il est nécessaire d'observer des règles précises pour contrôler la perte d'information

lors de la généralisation. Par ailleurs, plusieurs relations spatiales implicites entre des entités étudiées peuvent être envisagées. Une entité spatiale peut être située à côté d'une rivière et peut aussi appartenir à un quartier. Il est donc nécessaire d'effectuer un prétraitement afin de prendre en compte ces différentes proximités spatiales.

Dans cette étape, nous proposons d'explorer les données de deux manières différentes dans le but de regrouper des entités spatiales et de construire des zones géographiques homogènes pertinentes afin de gérer : (a) l'appartenance des entités spatiales à une région géographique construite par les êtres humaines ou des entités spatiales suivant un accident géographique (une montagne, une vallée, une rivière, etc.) ; et (b) la proximité à partir de la localisation des entités spatiales exprimée par leurs coordonnées Lambert (système de coordonnées géo-référencées).

Ces deux regroupements permettent d'étudier comment les évènements proches peuvent avoir des répercussions sur les entités étudiées, par exemple, dans la Figure 3.5, nous observons que le phénomène *tempête* peut affecter les entités spatiales A et B situées à côté de la rivière ainsi que des entités spatiales C et D placés sur une zone contiguë dans la vallée.

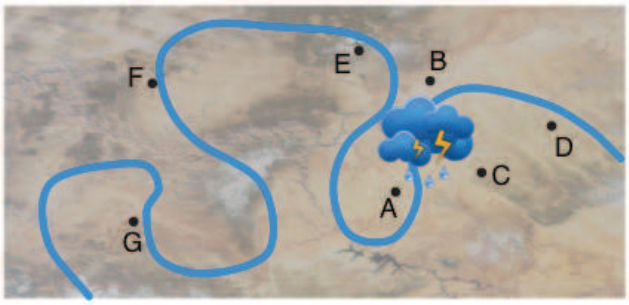


FIGURE 3.5 – Impact d'un phénomène spatio-temporel sur des entités spatiales

Enfin, le regroupement d'entités suivant ces deux critères n'obéit pas seulement à une classification descriptive des entités spatiales mais introduit le voisinage et la continuité (i.e., des entités qui suivent un cours d'eau) pour aboutir à la définition de nouvelles zones géographiques. Dans cette thèse, deux découpages de l'espace ont été utilisés :

– *l'appartenance* : deux entités sont regroupées si elles appartiennent à une même entité à un niveau de granularité spatiale plus élevé. Ce type de groupement appelé ω-appartenance permet d'étudier un ensemble d'entités qui appartiennent à une autre entité spatiale ou une configuration géographique notée par ω. Par exemple, nous pouvons regrouper toutes les maisons qui se trouvent à côté d'une rivière ou les maisons qui appartient à un quartier (ω est une rivière ou un quartier respectivement). Dans la Figure 3.6, le phénomène peut affecter les entités spatiales A et B qui se trouvent le long de la rivière, donc, ces entités spatiales sont considérés comme une seule zone et leurs données sont agrégées (e.g. moyenne des températures) ;

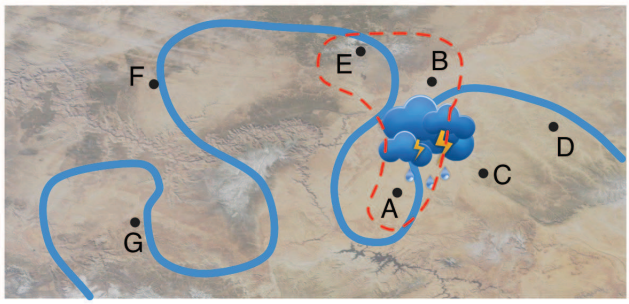


FIGURE 3.6 – Division de l'espace par la méthode ω-appartenance

– *l'agrégation* : l'espace est divisé en zones regroupant des entités spatiales selon leurs coordonnées Lambert. Dans chacune de ces zones, les entités spatiales situées dans une superficie couvrant ϵ km^2 sont regroupées. Par exemple, dans la Figure 3.7, les entités spatiales A, B et C se trouvant dans le carré de ϵ km^2 sont considérées comme une seule zone. Cette agrégation d'entités spatiales appelée ϵ-agrégation permet d'étudier l'impact d'un phénomène dans des zones suffisamment grandes pour observer des effets potentiellement non locaux. Ce type de groupement d'entités spatiales à été largement discuté par Goodchild et Zhang [2002].

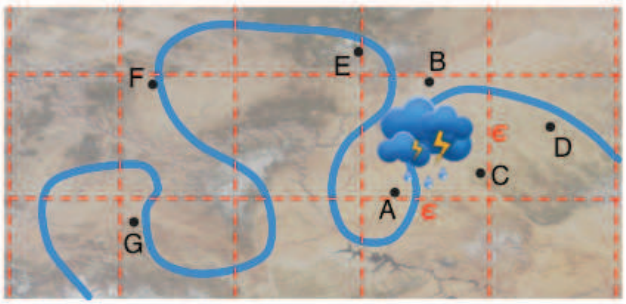


FIGURE 3.7 – Division de l'espace par la méthode ϵ-agrégation

3.2.4 Définitions et cadre formel

Comme discuté dans la Section 1.3, une base de données spatio-temporelles est définie comme un triplet $stDB = (D_T, D_S, D_A)$ où D_T est la dimension temporelle, D_S est la dimension spatiale et $D_A = \{D_{A_1}, D_{A_2}, \ldots, D_{A_p}\}$ est un ensemble des dimensions d'analyse associées aux attributs. Nous rappelons que la *dimension temporelle* est associée à un domaine de valeurs noté $dom(D_T) = \{T_1, T_2, \ldots, T_t\}$ où $\forall i \in [1..t]$, T_i est souvent appelé *estampille temporelle* et $T_1 < T_2 < \ldots < T_t$. Chaque dimension D_{A_i} $(\forall i \in [1..p])$ appartenant

à la *dimension d'analyse* est associée à un domaine de valeurs noté par $dom(A_i)$. Dans ce domaine, les valeurs peuvent être ordonnées ou non. La *dimension spatiale* est associée à un domaine de valeurs noté $dom(D_S) = \{Z_1, Z_2, \ldots Z_l\}$ où $\forall i \in [1..l]$, Z_i est une *zone*.

Définition 3.1 ***Item et Itemset***

Soit un item I*, une valeur littérale pour la dimension* D_{A_i}, $I \in dom(D_{A_i})$. *Un itemset,* $IS = (I_1 I_2 \ldots I_n)$ *avec* $n \leq p$ *est un ensemble non vide d'items tel que* $\forall i, j \in [1..n], \exists k, k' \in [1..p]$, $I_i \in dom(D_{A_k})$, $I_j \in dom(A_{D_{A_{k'}}})$ *et* $k \neq k'$.

Tous les items appartenant à un itemset sont associés à différentes dimensions d'analyse. Un itemset avec k items est appelé k-itemset.

Définition 3.2 ***Séquence d'itemsets***

Une séquence d'itemsets s *est une liste ordonnée dans le temps, non vide, d'itemsets notée* $\langle IS_1 IS_2 \ldots IS_p \rangle$ *où* IS_j *est un itemset.*

Une n-séquence est une séquence composée de n items. Par exemple, considérons les évènements produits dans la zone Z_1 du 12/03/2013 au 14/03/2013 selon la séquence $s = \langle (T_b H_m P_m)(T_m H_m P_b)(T_b H_m P_m 55) \rangle$ indiquée dans le Tableau 1.1. Cette séquence signifie que les ensembles d'évènements $(T_b H_m P_m)$, $(T_m H_m P_b)$ et $(T_b H_m P_m 55)$ se sont produits dans la même zone à trois dates successives. Dans cet exemple, s est une 10-séquence.

Définition 3.3 ***Sous-séquence***

Une séquence $\langle IS_1 IS_2 \ldots IS_p \rangle$ *est une sous-séquence d'une autre séquence* $\langle IS'_1 IS'_2 \ldots IS'_m \rangle$ *s'il existe des entiers* $i_1 < i_2 < \ldots < i_j < \ldots < i_p$ *tels que* $IS_1 \subseteq IS'_{i_1}, IS_2 \subseteq IS'_{i_2}, \ldots, IS_p \subseteq IS'_{i_p}$.

Par exemple, la séquence $s' = \langle (T_m H_m)(T_b H_m P_m 55) \rangle$ est une sous-séquence de s car $(T_m H_m) \subseteq (T_m H_m P_b)$ et $(T_b H_m P_m 55) \subseteq (T_b H_m P_m 55)$. Toutefois, $s'' = \langle (T_m H_h)(T_b H_m P_m 75) \rangle$ n'est pas une sous-séquence de s car les deux itemsets de s'' ne sont pas inclus dans trois itemsets de s.

Tous les évènements produits dans une même zone sont regroupés et triés par date. Ils constituent la séquence de données de la zone. Une zone supporte une séquence s si s est incluse dans la séquence de données de cette zone (s est une sous-séquence de la séquence de données).

Définition 3.4 ***Support d'une séquence***

Le support *d'une séquence* s *est alors calculé comme étant le pourcentage des séquences des données incluant* s. *Soit* $\mathtt{minSupp}$ *le support minimum fixé par l'utilisateur, une séquence qui vérifie le support minimum (i.e., dont le support est supérieur à* $\mathtt{minSupp}$*) est une* séquence fréquente.

Pour illustrer la mesure proposée, nous allons utiliser une base de séquences. Nous transformons la base de données spatio-temporelles présentée dans le Tableau 1.1 en une base de séquences présentée dans le Tableau 3.1 en regroupant des évènements appartenant à la dimension d'analyse (D_A) par jour et par zone. Ces séquences sont stockées dans une base appelée *base de séquences (seqDB)*.

TABLE 3.1 – Représentation des entités spatiales par des séquences

Zones	**Séquences**
Z_1	$s_1 = \langle(T_b H_m P_m)(T_m H_m P_b)(T_b H_m P_m 55)\rangle$
Z_2	$s_2 = \langle(T_m H_m P_m)(T_b H_m P_b)(T_b H_b P_m)\rangle$
Z_3	$s_3 = \langle(T_b H_m P_h 75)(T_m H_h P_b)(T_m H_h P_h 55)\rangle$

Chaque zone $Z_i \in dom(D_S)$ peut être représentée par sa séquence s_i, par exemple, la séquence $s_1 = \langle(T_b H_m P_m)(T_m H_m P_b)(T_b H_m P_m 55)\rangle$ représentent les évènement apparus dans la zone Z_1.

Le problème de la recherche de motifs séquentiels dans une base de données consiste à trouver les séquences dont le support est supérieur au support minimum spécifié, noté *minSupp*. Chacune de ces séquences fréquentes est communément appelée *motif séquentiel*.

Par exemple, la séquence $\langle(H_m P_m)(T_b)\rangle$ apparait dans deux séquences du Tableau 3.1 (s_1 et s_2). Son support est donc 2/3. Si le support minimum *minSupp* est fixé à 1, alors, cette séquence ne sera pas extraite.

Grâce aux deux méthodes de spatialisation décrites dans la Section 3.2.3, nous sommes capables de rapprocher les entités spatiales dans des zones et ainsi, construire la base de données de séquences utilisée lors de la phase de fouille de données. Le calcul ne se fera plus en fonction des stations mais des zones ainsi définies. Ceci permet d'obtenir des motifs séquentiels plus pertinents vis-à-vis du caractère hétérogène des données. Il faut préciser que l'on cherche à caractériser l'arrangement géographique des entités spatiales en explicitant les principes de leur distribution, indépendamment d'un attribut descriptif particulier : *les méthodes présentées ici ne prennent en compte que la localisation des entités étudiées*.

À cet égard, nous avons pré-traité les données afin de découper l'espace en zones. Les séquences de données sont ensuite obtenues en regroupant les données d'une même zone et en les triant par date. Ainsi, nous pouvons utiliser un algorithme classique d'extraction de motifs séquentiels, comme décrit ci-après.

3.2.5 Extraction des séquences spatialement fréquentes

Le problème de la recherche de motifs séquentiels a été introduit par Agrawal et Srikant [1995] dans le contexte du panier de la ménagère et appliqué avec succès dans de nombreux domaines comme la biologie [Wang *et al.*, 2004b; Salle *et al.*, 2009], la fouille d'usage du Web [Pei *et al.*, 2000; Masseglia *et al.*, 2008], la détection d'anomalies [Rabatel *et al.*, 2010], la fouille de flux de données [Marascu et Masseglia, 2006] ou la description des comportements au sein d'un groupe [Perera *et al.*, 2009]. D'autres approches [Julea *et al.*, 2008] utilisent les motifs séquentiels pour décrire les évolutions temporelles des pixels au sein d'une série temporelle d'images satellites.

Comme indiqué précédemment, le principal défi de la recherche de motifs séquentiels à partir d'une base de données est d'extraire des séquences pour lesquelles le support est supérieur ou égal à un seuil minimal *minsupp*.

Pour extraire les motifs séquentiels, l'algorithme PrefixSpan [Mortazavi-Asl *et al.*, 2000] a été adopté en raison de son efficacité sur des grands volumes de données. Cette méthode utilise la stratégie *diviser pour régner* en effectuant une *recherche en profondeur* de l'espace de recherche avec des projections successives de la base de données.

Dans les paragraphes suivants, nous présentons quelques définitions pour mieux comprendre l'algorithme PrefixSpan. Nous supposons que tous les items sont triés par ordre alphabétique.

Définition 3.5 ***Préfixe d'une séquence***
Nous définissons la fonction $\mathtt{préfixe} : S \times \mathbb{N} \rightarrow S$ *où* S *est un ensemble de séquences,* $\mathbb{N}$ *est un ensemble de nombres entiers positifs et* $\mathtt{préfixe}(s,k) = s[1:k]$. *En d'autres termes,* $\mathtt{préfixe}(s,k)$ *retourne les premiers* k *items, i.e., le* préfixe *de* s.

Définition 3.6 ***Suffixe d'une séquence***
Nous définissons la fonction $\mathtt{suffixe} : S \times S \rightarrow S$ *tel que* $\mathtt{suffixe}(s,s') = s[m+1:n]$ *si et seulement si,* s' *est un préfixe de* s *avec* m *items et* s *est une séquence comportant* n *items.*

Par exemple, soient les séquence $s_1 = \langle(T_m H_b P_m)(T_h)\rangle$ et $s_2 = \langle(T_m H_b)\rangle$, le suffixe de s_1 comparé au préfixe s_2 est $\langle(\_P_m)(T_h)\rangle$.

Définition 3.7 ***Projection d'une base de séquences***
Soit s *une séquence dans la base de séquences* $\mathtt{seqDB}$. *La base de données* s*-projetée, notée* $\mathtt{seqDB}|_s$ *est un ensemble de suffixes de séquences de* $\mathtt{seqDB}$ *préfixées par* s.

L'algorithme PrefixSpan

Nous présentons un pseudo-code simplifié de l'algorithme PrefixSpan. Il prend en entrée une séquence s et un seuil minimal *minsupp* et retourne toutes les séquences fréquentes préfixées par s, i.e., celles ayant un support supérieur à *minSupp*. L'Algorithme 1 présente le pseudo-code de PrefixSpan.

Algorithme 1: $PrefixSpan(\alpha, seqDB|_\alpha)$

ENTRÉ: un préfixe α et la base α-projétée $seqDB|_\alpha$

1. trouver l'item x tel que $support_{seqDB|_\alpha}(x) \leq minsupp$
2. ajouter x à α pour étendre le motif séquentiel αx et le garder
3. construire la base αx-projetée $(seqDB|_\alpha)|_x$ pour chaque αx et appeler $PrefixSpan(\alpha x, (seqDB|_\alpha)|_x)$

Les motifs fréquents sont progressivement étendus au cours d'une exploration en profondeur de l'espace de recherche. D'abord, l'ensemble des items fréquents x sont extraits de la base de données projetée $seqDB|_\alpha$. Il faut noter que, dans le premier appel récursif, $seqDB|_\alpha$ correspond à la base de données initiale $seqDB$ (car α={}).

Ensuite, pour chaque item x, l'algorithme étend le motif séquentiel α avec x. Deux extensions sont possibles : (1) en ajoutant x au dernier itemset de α, soit $(x\alpha)$; ou (2) en insérant x après le dernier itemset de α, soit $(x)(\alpha)$ (i.e., la prochaine estampille temporelle). Le support pour ces deux motifs est calculé et seul les motifs fréquents sont conservés.

Finalement, pour chaque motif fréquent, l'algorithme effectue une autre projection de la base de données en utilisant $seqDB|_\alpha$ et étend de manière récursive le motif en invoquant à nouveau la fonction *PrefixSpan*. L'algorithme s'arrête lorsque plus aucune projection ne peut être générée.

Comme nous le montrerons dans la Section 5.3, cette méthode a été appliquée avec succès sur deux jeux de données. Si les motifs obtenus sont très intéressants pour les experts, ils ne mettent pas en valeur la propagation des évènements d'une zone à l'autre. Pour résoudre ce problème, nous proposons un nouveau type de motifs dans la Section 3.3 : les motifs spatio-séquentiels.

3.3 Motifs spatio-séquentiels

La deuxième approche étudiée consiste à introduire des caractéristiques spatiales dans les motifs recherchés, c'est-à-dire en modifiant un algorithme de recherche de motifs séquentiels pour prendre en compte les relations spatiales décrites dans la Section 3.1.

Dans cette approche, nous étudions l'extraction de séquences représentant l'évolution d'une zone en fonction de ses caractéristiques et de celles de son voisinage. Pour cela, nous avons défini un nouveau concept de séquences spatio-temporelles d'itemsets spatiaux (i.e., des séquences d'ensembles) appelée S2P (Spatio-Sequential Patterns). Notre approche étend donc les travaux de Tsoukatos et Gunopulos [2001] en intégrant les interactions entre la zone d'étude et son environnement proche. À titre d'illustration, grâce à l'approche citée auparavant, nous pouvons extraire des informations du type *dans la ville de Nîmes, de fortes pluies se sont produites après une augmentation considérable de l'humidité*. A contrario, grâce à notre deuxième approche, des informations seront extraites du

type : *dans la ville de Nîmes, le pourcentage d'humidité a augmenté considérablement car auparavant il a eu de fortes pluies à Nîmes et des hautes températures dans les villes voisines.* Les approches existantes ne permettent pas d'extraire ce type d'information.

3.3.1 Préliminaires

Nous allons étendre les définitions présentées dans la Section 3.2.4 de façon à prendre en compte la dynamique spatiale des données.

Pour cela, nous définissons la relation $dans$ entre une zone Z et un itemset IS comme l'occurrence de l'itemset IS dans la zone Z au temps t dans la base de données $stDB$. Plus formellement :

$$\begin{cases} dans(IS,Z,t) = vrai \text{ si } IS \text{ apparait dans } stDB \text{ pour la zone } Z \text{ au temps } t \\ dans(IS,Z,t) = faux \text{ sinon} \end{cases} \tag{3.1}$$

Maintenant, nous définissons la notion de *voisin* entre zones. Différentes relations de voisinage peuvent être définies sur des zones. Deux zones sont voisines [3] si :

$$\begin{cases} voisin(Z_i,Z_j) = vrai \text{ si } Z_i \text{ et } Z_j \text{ sont voisines} \\ voisin(Z_i,Z_j) = faux \text{ sinon} \end{cases} \tag{3.2}$$

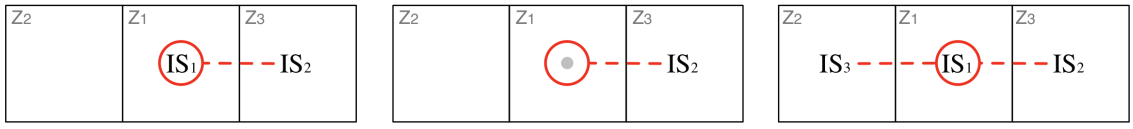
Définition 3.8 ***Itemset spatial***

Soient IS_p *and* IS_q *deux itemsets,* IS_p *et* IS_q *sont spatialement proches si* $\exists Z_i, Z_j \in dom(D_S)$ *et* $\exists t \in dom(D_T)$ *tel que* $dans(IS_p,Z_i,t) \wedge dans(IS_q,Z_j,t) \wedge voisin(Z_i,Z_j)$ *est vrai. Deux itemsets* IS_p *et* IS_q *qui sont spatialement proches, forment un itemset spatial noté* $I_{ST} = IS_p \cdot IS_q$.

Pour alléger les notations, nous introduisons un opérateur de groupement d'itemsets associé à l'opérateur $\cdot$ (*voisin*) et noté *[]*. Le symbole θ représente l'absence d'itemsets dans une zone. La Figure 3.8 montre les trois types d'itemsets spatiaux que nous pouvons construire en utilisant ces opérateurs. Les lignes pointillées représentent le voisinage spatial.

Par exemple, l'itemset spatial $I_{ST} = (T_m \cdot H_m P_b)$ signifie que les évènements T_m et $H_m P_b$ apparaissent dans une zone voisine au même temps. L'itemset spatial $I_{ST} = (\theta \cdot [T_m; H_b P_m])$ indique que T_m et $H_b P_m$ apparaissent dans deux zones voisines et qu'aucun évènement n'est apparu dans la zone d'étude.

3. Ici la notion de voisinage, au sens commun du terme, peut être définie comme des zones qui partagent une frontière ou des zones qui sont proches. Cette notion de voisinage est liée à la disposition des entités spatiales dans l'espace.



(a) (b) (c)

FIGURE 3.8 – Représentation graphique des itemset spatiaux (a) $IS_1 \cdot IS_2$ (b) $\theta \cdot IS_2$ (c) $IS_1 \cdot [IS_2; IS_3]$

Définition 3.9 ***Association entre zone, itemset spatial et le temps***

Soit $I_{ST} = IS_p \cdot IS_q$ *un itemset spatial,* $Z \in dom(D_S)$ *une zone et* $t \in dom(D_T)$ *une estampille temporelle, nous définissons la relation* vérifier *qui représente la présence de l'itemset spatial* I_{ST} *dans* Z *au temps* t *comme suit :*

$$\begin{cases} vérifier(I_{ST}, Z, t) = vrai \textit{ si } dans(IS_p, Z, t) = vrai \textit{ et } \exists Z' \in dom(D_S) \textit{ tel que} \\ \qquad\qquad\qquad\qquad voisin(Z, Z') = vrai \textit{ et } dans(IS_q, Z', t) = vrai \\ vérifier(I_{ST}, Z, t) = faux \textit{ sinon} \end{cases} \tag{3.3}$$

Définition 3.10 ***Inclusion d'itemsets spatiaux***

Un itemset spatial $I_{ST} = IS_p \cdot IS_q$ *est inclus, noté par* $\subseteq$*, dans autre itemset spatial* $I'_{ST} = IS'_k \cdot IS'_l$*, si et seulement si* $IS_p \subseteq IS'_k$ *et* $IS_q \subseteq IS'_l$.

Par exemple, l'itemset spatial $I_{ST} = (T_b H_m \cdot P_h)$ est inclus dans l'itemset spatial $I'_{ST} = (T_b H_m \cdot P_h 75)$ car $(T_b H_m) \subseteq (T_b H_m)$ et $(P_h) \subseteq (P_h 75)$.

Nous modélisons la notion *d'évolution* d'évènements dans les zones en prenant en compte leur relation de voisinage via la notion de séquence spatiale.

Définition 3.11 ***Séquence spatiale***

Une séquence spatiale ou simplement 2S est une liste ordonnée d'itemsets spatiaux notée par $s = \langle I_{ST_1} I_{ST_2} \dots I_{ST_m} \rangle$ *où* I_{ST_i}, $I_{ST_{i+1}}$ *satisfaisant la contrainte de séquentialité temporelle, i.e.,* $i \in [1..m-1]$.

La 2S $s = \langle (T_b)(\theta \cdot [H_m; P_h])(P_m \cdot [H_b; T_m]) \rangle$ est illustrée dans la Figure 3.9 pour la zone Z_1 où les flèches représentent la dynamique temporelle et les lignes pointillées représentent le voisinage proche.

Une relation de généralisation/spécialisation entre deux 2S peut être définie de la manière suivante :

Définition 3.12 ***Inclusion de séquences spatiales (2S)***

Une 2S notée $s = \langle I_{ST_1} I_{ST_2} \dots I_{ST_m} \rangle$ *est plus spécifique (incluse) qu'une autre 2S* $s' = \langle I'_{ST_1} I'_{ST_2} \dots I'_{ST_n} \rangle$*, notée par* $s \preceq s'$*, s'il existe* $j_1 \leq \dots \leq j_m$ *tel que* $I_{ST_1} \subseteq I'_{ST_{j_1}}, I_{ST_2} \subseteq I'_{ST_{j_2}}, \dots, I_{ST_m} \subseteq I'_{ST_{j_m}}$.

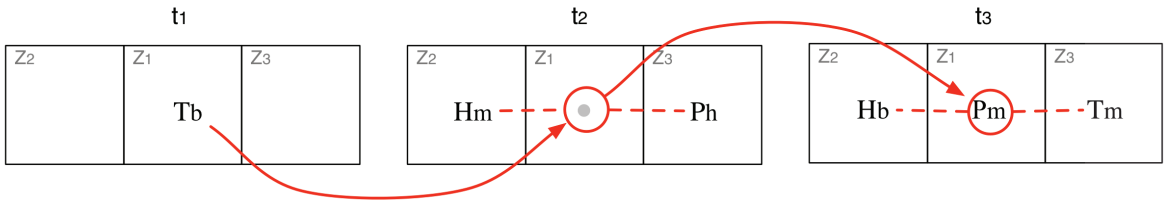


FIGURE 3.9 – Dynamique spatio-temporelle du motif $\langle(T_b)(\theta \cdot [H_m; P_h])(P_m \cdot [H_b; T_m])\rangle$

Par exemple, la 2S $s = \langle(T_b H_m \cdot H_b P_h)(55)\rangle$ est plus spécifique (incluse) dans la 2S $s' = \langle(T_b P_m \cdot H_b P_h)(55 \cdot P_h)\rangle$ car $(T_b H_m \cdot H_b P_h) \subseteq (T_b H_m \cdot H_b P_h)$ et $(55) \subseteq (55 \cdot P_h)$.

Définition 3.13 ***Préfixe de séquences spatiales (2S)***

Nous allons étendre la Définition 3.5 aux séquences spatiales. La fonction $\mathtt{préfixe}(s,k) = s[1:k]$ *renvoi les premiers* k *items, i.e., le préfixe de* s *où* s *est une 2S et* k *es un nombre entier positif.*

Définition 3.14 ***Suffixe de séquences spatiales (2S)***

Nous allons étendre la Définition 3.6 aux séquences spatiales. Le $\mathtt{suffixe}(s,s') = s[m+1:n]$ *renvoie les 2S contenant les* $(n-m)$ *derniers items de la 2S* s *préfixée par la 2S* s'.

Par exemple, soient les 2S $s_1 = \langle(T_b H_m \cdot H_b P_h)(55)\rangle$ et $s_2 = \langle(T_b H_m)\rangle$, le suffixe s_1 comparé au préfixe s_2 est $\langle(\_ \cdot H_b P_h)(55)\rangle$.

Définition 3.15 ***Projection d'une base de séquences***

Soit s *une 2S présente dans la base de séquences* $\mathtt{seqDB}$. *La base de données* s*-projetée, notée par* $\mathtt{seqDB}|_s$ *est l'ensemble des suffixes de la séquence spatiale dans* $\mathtt{seqDB}$ *préfixée par* s.

Par exemple, soit la base de séquences $\mathtt{seqDB}$ contenant les 2S $s_1 = \langle(T_b H_m \cdot H_b P_h)(55)\rangle$ et $s_2 = \langle(T_b H_m)(T_b \cdot P_m)\rangle$. Soit la 2S $s_3 = \langle(T_b H_m)\rangle$. La projection de la base de séquences $\mathtt{seqDB}$ comparée au préfixe s_3 (représentée par $\mathtt{seqDB}|_{s_3}$) est $\langle(\_ \cdot H_b P_h)(55)\rangle$ et $\langle(T_b \cdot P_m)\rangle$.

De manière générale, les algorithmes de fouille de données doivent faire face à une énorme complexité qui est associée essentiellement à la taille de la base de données à fouiller. Dans notre approche, cette complexité peut augmenter considérablement à cause de la relation de voisinage qui est présente au moment du processus de la fouille. Il est donc nécessaire de trouver des heuristiques pour élaguer l'espace de recherche. Ces heuristiques peuvent être de nature syntaxique (e.g. ne s'intéresser qu'à un sous-ensemble des motifs) et/ou basées sur des métriques statistiques (e.g. élaguer tous les motifs dont le support est inférieur à un seuil minimal).

Dans la suite de cette thèse, nous proposons deux mesures d'élagage basées sur des métriques statistiques adaptées aux séquences spatiales.

3.3.2 Mesures d'élagage pour des séquences spatiales 2S

Comme cela a été discuté dans [Joshi *et al.*, 2001], l'une des tâches les plus importantes dans le processus de fouille de données est de savoir comment la fréquence d'apparition d'une séquence est estimée. Dans la littérature, il existe différentes mesures pour évaluer les motifs extraits lors de l'étape de fouille de données (pour un aperçu, voir [McGarry, 2005; Geng et Hamilton, 2006]). Par exemple, Mohan *et al.* [2012], proposent la mesure appelée *indice de participation en cascade* (*IPC* en abrégé), qui est définie comme le nombre minimal d'occurrences d'un type d'évènement participant à un motif divisé par le nombre d'instances du type d'évènement dans toute la base de données. Cette mesure est composée par deux indices : l'indice de participation spatiale et l'indice de participation temporelle. Pour estimer l'IPC, les auteurs comptent "globalement" les occurrences d'un motif dans la base de données.

Dans cette thèse, nous introduisons deux mesures d'élagage pour filtrer les séquences spatiales :

1. La première est une adaptation du *support* utilisé pour l'extraction de motifs séquentiels ([Agrawal et Srikant, 1995]).
2. La deuxième est une adaptation de l'*indice de participation en cascade* utilisé dans [Mohan *et al.*, 2012]. Cette mesure a été adaptée en prenant en compte deux aspects : (1) tout d'abord, dans l'*indice de participation temporelle*, nous proposons d'évaluer les occurrences d'une séquence spatiale $2S$ "localement" (par rapport à l'échelle globale), en comptant le nombre d'occurrences d'évènements dans une séquence où le motif apparaît (par rapport à la base de données entière) ; et (2) comme le détaille Wu *et al.* [2010], les séquences nulles [4] sont extrêmement importantes pour le calcul d'une mesure d'élagage. Notre proposition tient compte de cette contrainte en évaluant uniquement les séquences contenant la séquence spatiale $2S$ étudiée.

Dans la suite de cette section, nous allons détailler ces deux mesures.

Support d'un motif spatio-séquentiel

Dans la littérature portant sur la fouille de séquences fréquentes, un exemple classique est celui du *panier de la ménagère* où la base de données est composée d'un ensemble de séquences correspondant à des transactions faites par des clients. Chaque transaction est

4. Une séquence nulle, par rapport à un item I, où $I \in dom(A_i)$, est une séquence contenue dans la base de séquences qui ne contienne pas l'item I.

constituée d'un ID client, de la date de la transaction et des articles achetés dans les transactions. Le support absolu d'une séquence s est définie comme le nombre de séquences dans la base de séquences qui contient la séquence s [Agrawal et Srikant, 1995].

La base de données spatio-temporelles est semblable à une base de séquences (transactions), car les informations d'une zone à différents moments peuvent être considérées comme une séquence. La principale différence est la relation de voisinage associée à la dimension spatiale. En conséquence, nous définissons un nouveau *support absolu* pour des séquences spatiales défini comme le nombre de zones contenant la séquence étudiée et satisfaisant les contraintes de proximité des itemsets spatiaux. Plus formellement, cette mesure peut être définie comme suit :

Définition 3.16 ***Support absolu d'une séquence spatiale***

Soit la 2S $s = \langle I_{ST_1} I_{ST_2} \dots I_{ST_n} \rangle$*, le support absolu de* s *représenté par* $supp_{abs}(s, seqDB)$ *ou simplement* $supp_{abs}(s)$ *est défini comme le nombre de séquences de la base* seqBD *qui vérifient* s*. Autrement dit :*

$$supp_{abs}(s, seqBD) = supp_{abs}(s) = |\{Z \in dom(D_S) | \forall i \in [1..n], \exists T_i \in dom(D_T), \text{ et } vérifier(I_{ST_i}, Z, T_i) = vrai\}| \quad (3.4)$$

De la même manière, nous définissons le *support relatif* pour une séquence 2S comme le ratio entre le nombre de zones qui vérifient la séquence 2S et le nombre de zones total.

Définition 3.17 ***Support relatif d'une séquence spatiale***

Soit la séquence $s = \langle I_{ST_1} I_{ST_2} \dots I_{ST_n} \rangle$*, le support relatif de* s *noté par* $supp_{rel}(s, seqBD)$ *ou plus simplement* $supp(s)$*, est définie comme :*

$$supp_{rel}(s, seqBD) = supp(s) = \frac{supp_{abs}(s)}{|dom(D_S)|} \quad (3.5)$$

Par exemple, soit la relation de voisinage montrée dans la Figure 3.11 et les séquences présentées dans le Tableau 3.1. Chaque séquence représente l'évolution d'un ensemble d'évènements par zone (cf., Table 1.1).

Soit une séquence spatiale $s = \langle (H_m)(T_b \cdot T_m) \rangle$, le support relatif de s est 2/3. La Figure 3.10 illustre la dynamique de s. Les flèches représentent la dynamique temporelle tandis que les lignes pointillées représentent la relation de voisinage.

Définition 3.18 ***Motif spatio-séquentiel***

Soient la 2S s *et* minSupp *le support minimum spécifié par l'utilisateur, si* $supp_{rel}(s, seqDB) \geq minSupp$ *alors* s *est une 2S fréquente appelée motif spatio-séquentiel.*

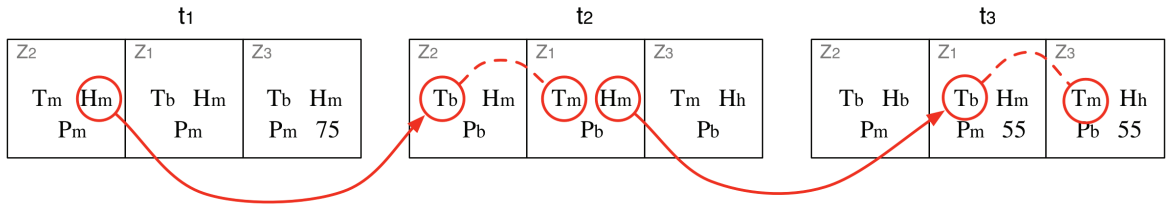


FIGURE 3.10 – Représentation graphique du calcul du support du motif $\langle (H_m)(T_b \cdot T_m) \rangle$

Indice de participation spatio-temporel

La mesure précédente ne tient pas compte de la participation d'un item dans un motif spatio-séquentiel. Il n'évalue pas les occurrences d'un motif apparaissant dans la même zone à des temps différents. Pour mettre en évidence ces aspects, nous proposons une adaptation de l'*indice de participation en cascade* défini par [Mohan *et al.*, 2012] qui est une combinaison de deux mesures : (1) *l'indice de participation spatial*, qui prend en compte le nombre de zones supportant le motif ; et (2) *l'indice de participation temporelle*, qui prend en compte le nombre d'apparitions d'un motif dans une zone à différents moments.

Pour présenter ces deux mesures, nous devons tout d'abord définir les ratios de participation spatial et temporel.

Définition 3.19 ***Ratio de participation spatial***

Soient s *une séquence spatiale et* I *un item de* s. *Le ratio de participation spatial de l'item* I *dans* s, *noté par* $SPr(s, I)$ *est le nombre de zones qui contiennent* s *divisé pour le nombre de zones contenant l'item* I :

$$SPr(s, I) = \frac{supp_{abs}(s)}{supp_{abs}(I)} \tag{3.6}$$

Définition 3.20 ***Indice de participation spatial***

Soit s *une séquence spatiale, l'indice de participation spatial de* s *noté par* $SPi(s)$ *est le minimum des ratios de participation spatiaux :*

$$SPi(s) = MIN_{\forall I \in dom(A_i), I \in s}\{SPr(s, I)\} \tag{3.7}$$

Définition 3.21 ***Ratio de participation temporelle***

Soit s *une séquence spatiale, soit* I *un item de* s *et soit* s_i *une séquence qui représente des évènements apparus dans une zone* Z_i *et qui supporte* s. *Le ratio de participation temporelle de l'item* I *dans* s *noté par* $TPr(s, I, s_i)$ *est le nombre d'instances de l'item* I *qui participent dans* s *divisé par le nombre total d'instances de l'item* I *dans la séquence* s_i *supportant* s :

$$TPr(s, I, s_i) = \frac{nbInstances(I) \text{ qui apparaissent dans } s}{nbInstances(I) \text{ dans toute la séquence } s_i} \tag{3.8}$$

Définition 3.22 ***Indice de participation temporelle***

Soit s *une séquence spatiale, l'indice de participation temporelle de* s *noté par* $TPi(s)$ *est le minimum des indices de participation temporelles calculés pour chaque séquence de la base de données :*

$$TPi(s) = MIN_{s_i \in seqBD, i \in dom(D_S)} \left\{ MIN_{I \in dom(A_i), I \in s} \{TPr(s, I, s_i)\} \right\} \tag{3.9}$$

Définition 3.23 ***Indice de participation spatio-temporel***

L'indice de participation spatio-temporel d'une séquence spatiale s *noté par* $STPi(s)$*, est le produit des deux mesures présentées auparavant :*

$$STPi(s) = SPi(s) \times TPi(s) \tag{3.10}$$

Par exemple, considérons la séquence spatiale présentée dans la Definition 3.17 i.e., $s = \langle (H_m)(T_b \cdot T_m) \rangle$, la relation de voisinage montrée dans la Figure 3.11 et la base de séquences représentée dans le Tableau 3.1. L'indice de participation spatio-temporel de s est $1/3$. Nous l'obtenons grâce au calcul ci-dessous :

Initialement, nous calculons l'indice de participation spatial

$$\begin{aligned} SPi &= MIN\{SPr(s, H_m), SPr(s, T_b), SPr(s, \cdot T_m)\} \\ &= MIN\left\{\frac{2}{3}, \frac{2}{3}, \frac{2}{3}\right\} = \frac{2}{3} \end{aligned}$$

Puis, l'indice de participation temporelle :

$$\begin{aligned} TPi =& MIN_{\forall s_i, i \in dom(D_S)} \{MIN\{TPr(s, H_m), TPr(s, T_b), TPr(s, \cdot T_m)\}\} \\ =& MIN\left\{\left\{\frac{2}{3}, \frac{1}{1}, \frac{2}{4}\right\}, \left\{\frac{1}{2}, \frac{1}{1}, \frac{1}{1}\right\}, \{\}\right\} \\ =& MIN\left\{\left\{\frac{1}{2}\right\}, \left\{\frac{1}{2}\right\}\right\} = \frac{1}{2} \end{aligned}$$

Finalement, les deux résultats obtenus précédemment sont remplacés dans la formule de la Définition 3.23 :

$$\begin{aligned} STPi(s) &= SPi(s) \times TPi(s) \\ &= \frac{2}{3} \times \frac{1}{2} = \frac{1}{3} \end{aligned}$$

Proposition 3.24 *(Anti-monotonie) :*
L'indice de participation spatio-temporel est anti-monotone [Agrawal et Srikant, 1994], i.e., monotoniquement non incrémental par rapport à la taille du motif.

Nous allons démontrer cette proposition indépendamment pour chaque composante - spatiale et temporelle - de la mesure.

Proposition 3.25 *(Anti-monotonie) :*
L'indice de participation spatial est anti-monotone.

Démonstration :
Le *ratio de participation spatial* est anti-monotone car un item $I \in dom(A_i)$ qui participe dans une séquence spatiale s participe aussi à la séquence spatiale s' où $s \subseteq s'$. L'*indice de participation spatial* est aussi anti-monotone car le *ratio de participation spatial* est anti-monotone et les inégalités suivantes sont conservées :

$$\begin{aligned} supp_{abs}(s') &\leq supp_{abs}(s) \\ \frac{supp_{abs}(s')}{supp_{abs}(I)} &\leq \frac{supp_{abs}(s)}{supp_{abs}(I)} \\ SPr(s', I) &\leq SPr(s, I) \\ MIN_{I \in dom(A_i), I \in s'}\{SPr(s', I))\} &\leq MIN_{I \in dom(A_i), I \in s}\{SPr(s, I))\} \end{aligned}$$

En consequence, l'*indice de participation spatial* es anti-monotone.

□

Proposition 3.26 *(Anti-monotonie) :*
L'indice de participation temporelle est anti-monotone.

Démonstration :
Soient s et s' deux séquences spatiales respectivement de taille $k-1$ et k où $s \subseteq s'$. Nous déterminons que $TPi(s') \leq TPi(s)$ par addition d'un item. Soit $I \in dom(A_i)$ un item où $I \in s'$ et $I \notin s$. Par définition du TPi, nous avons :

$$TPi(s') = MIN_{\forall s_i, i \in dom(D_S)} \{ MIN_{\forall I \in dom(A_i), I \in s} \{ TPi(s), TPr(s', I, s_i) \} \}$$
$$TPi(s') \leq TPi(s)$$

Alors, l'*indice de participation temporelle* est anti-monotone.

□

Finalement, si $s \subseteq s'$, alors, l'*indice de participation spatio-temporel* est anti-monotone, i.e.,

$$STPi(s') \leq STPi(s)$$

La propriété d'anti-monotonie est très importante dans le processus d'extraction de motifs spatio-séquentiels car elle permet d'élaguer l'espace de recherche. En effet, si le STPi d'une séquence est faible (inférieure à un seuil), aucun des STPi de ses super-séquences ne pourra dépasser ce seuil et il est inutile de les explorer. Nous montrerons dans la section suivante, comment cette propriété est utilisée.

3.3.3 Algorithmes d'extraction de motifs spatio-séquentiels

Dans cette section, nous proposons deux algorithmes associés aux deux stratégies les plus courantes en fouille de données pour rechercher des motifs fréquents dans une base de données. Dans un premier temps, nous utilisons la stratégie d'exploration en largeur ou par niveau *(BFS, breadth-first search)* et nous proposons l'algorithme appelé *BFS-S2PMiner*. Dans un deuxième temps, nous proposons l'algorithme appelé *DFS-S2PMiner* s'appuyant sur une stratégie d'exploration en profondeur basée sur des projections successives de la base de données *(DFS, depth-first search)*.

Approche par niveau

Dans les algorithmes de recherche par niveau (*générer-tester-élaguer*), la démarche consiste à extraire tout d'abord les items fréquents, puis, pour chaque itération k, à générer un ensemble de candidats à partir des motifs fréquents générés à l'itération $k-1$. À la fin de chaque itération, une phase d'élagage, basée sur une propriété d'anti-monotonie limite le nombre de motifs candidats générés. Tous les candidats sont testés et seuls ceux qui sont fréquents sont utilisés pour générer des motifs candidats de $k+1$ items lors de l'itération suivante. L'algorithme s'arrête lorsque l'ensemble des candidats est vide. Il existe de nombreux algorithmes basés sur cette approche, parmi lesquels nous trouvons *Apriori* [Agrawal et Srikant, 1994], *GSP* [Srikant et Agrawal, 1996] ou *SPADE* [Zaki, 2001]. Nous proposons dans cette section un nouvel algorithme dérivé de l'algorithme *SPADE* introduit par Zaki dans [Zaki, 2001]. L'Algorithme 2 correspond au pseudo-code de notre proposition.

TABLE 3.2 – Génération de motifs candidats

Séquence α	**Séquence β**	**Si**	**Motif candidat γ**
$\langle(\rho X)\rangle$	$\langle(\rho Y)\rangle$	$X < Y$	$\langle(\rho XY)\rangle$
$\langle(\rho X)\rangle$	$\langle(\rho)(Y)\rangle$		$\langle(\rho X)(Y)\rangle$
$\langle(\rho)(X)\rangle$	$\langle(\rho)(Y)\rangle$	$X < Y$	$\langle(\rho)(XY)\rangle$
$\langle(\rho X)\rangle$	$\langle(\rho \cdot Y)\rangle$		$\langle(\rho X \cdot Y)\rangle$
$\langle(\rho X)\rangle$	$\langle(\rho)(\theta \cdot Y)\rangle$		$\langle(\rho X)(\theta \cdot Y)\rangle$
$\langle(\rho \cdot X)\rangle$	$\langle(\rho)(Y)\rangle$		$\langle(\rho \cdot X)(Y)\rangle$
$\langle(\rho \cdot X)\rangle$	$\langle(\rho \cdot Y)\rangle$	$X < Y$	$\langle(\rho \cdot [X;Y])\rangle$
$\langle(\rho \cdot XY)\rangle$	$\langle(\rho \cdot Z)\rangle$	$X < Z$ et $Y < Z$	$\langle(\rho \cdot [XY;Z])\rangle$
$\langle(\rho \cdot X)\rangle$	$\langle(\rho \cdot YZ)\rangle$	$X < Y$ et $X < Z$	$\langle(\rho \cdot [X;YZ])\rangle$
$\langle(\rho)(\theta \cdot X)\rangle$	$\langle(\rho)(\theta \cdot Y)\rangle$	$X < Y$	$\langle(\rho)(\theta \cdot [X;Y])\rangle$

Dans un premier temps, notre algorithme extrait les items fréquents (lignes 1 à 8 de l'Algorithme 2) pour construire l'ensemble F_1. À partir de cet ensemble F_1, nous construisons les candidats de l'itération 2. De façon générale, à chaque niveau, nous construisons l'ensemble des motifs de $k+1$ items à partir de ceux comportant k items. Notons qu'à chaque itération k, les motifs générés ont k items. Par exemple, considérons deux motifs spatio-séquentiels $\langle(XY \cdot Z)\rangle$ et $\langle(XY)(Y)\rangle$, tous les deux composés de trois items. L'itemset commun à ces deux motifs est $\langle(XY)\rangle$ donc, nous pouvons générer le motif $\langle(XY \cdot Z)(Y)\rangle$ comportant quatre items.

Plus formellement, soient les séquences α et β ayant le même préfixe ρ. Dans l'étape de génération de candidats, nous construisons des motifs candidats γ ayant le même préfixe conformément aux règles montrées dans le Tableau 3.2.

Dans l'étape d'élagage, l'algorithme limite les motifs candidats à ceux pour lesquels tous les sous-motifs construits à l'itération $k-1$ sont fréquents (ligne 15). Par exemple, le motif $\langle(XYZ)\rangle$ sera candidat si et seulement si ses sous-motifs $\langle(XY)\rangle$, $\langle(XZ)\rangle$ et $\langle(YZ)\rangle$ sont fréquents.

Finalement, une fois l'élagage effectué, l'algorithme teste chacun des motifs candidats (ligne 16 de l'Algorithme 2) et utilise les motifs spatio-séquentiels découverts pour débuter l'itération suivante.

Pour la construction des motifs spatio-séquentiels, nous utilisons, dans l'Algorithme 2, un tableau comme structure de données pour stocker la base de données spatio-temporelles sous la forme de séquences. Ce tableau (voir l'exemple du Tableau 3.4) est divisé en deux parties : la première stocke les séquences qui représentent l'évolution dans le temps de chaque zone de la base de données spatio-temporelles et la deuxième partie stocke les séquences associées à l'évolution dans le temps de leurs zones voisines.

Exemple : Nous allons illustrer notre proposition à partir de la base de séquences seqDB (Tableau 3.3) construite à partir de la base de données spatio-temporelles décrite dans le

Algorithme 2: BFS-S2PMiner

ENTRÉ: Une base de séquences $seqBD$, une relation de voisinage L
et un support minimum $minSupp$
SORTIE: L'ensemble de motifs spatio-séquentiels

1. **pour tout** $i \subseteq dom(A_i)$ où $i \in [1..p]$ **faire**
2. **si** $supp(i) \geq minSupp$ **alors**
3. $F_1 \leftarrow F_1 \cup \{i\}$
4. **finsi**
5. **si** $supp(\cdot i) \geq minSupp$ **alors**
6. $F_1 \leftarrow F_1 \cup \{\cdot i\}$
7. **finsi**
8. **fin pour**
9. $k \leftarrow 1$;
10. **tant que** $F_k \neq \emptyset$ **faire**
11. $F_{k+1} \leftarrow \emptyset$;
12. **pour tout** $\alpha \in F_k, \beta \in F_k$ **faire**
13. **si** *prefix*$(\alpha) =$ *prefix*(β) **alors**
14. $\gamma \leftarrow union(\alpha, \beta)$;
15. **si** $AllSubSeqFreq(F_k, \gamma)$ **alors**
16. **si** $supp(\gamma) > supp\_min$ **alors**
17. $F_{k+1} \leftarrow F_{k+1} \cup \{\gamma\}$;
18. **finsi**
19. **finsi**
20. **finsi**
21. **fin pour**
22. $k \leftarrow k+1$;
23. **fin tant que**
24. **retourner** $F_1 \cup F_2 \cup \ldots \cup F_k$;

Tableau 1.1, de la relation de voisinage décrite dans la Figure 3.11 ainsi que d'un support minimum *minSupp* égal à 2/3.

TABLE 3.3 – Exemple de une base de séquences $seqBD$

Séquences
$s_1 = \langle (T_b H_m P_m)(T_m H_m P_b)(T_b H_m P_m 55) \rangle$
$s_2 = \langle (T_m H_m P_m)(T_b H_m P_b)(T_b H_b P_m) \rangle$
$s_3 = \langle (T_b H_m P_h 75)(T_m H_h P_b)(T_m H_h P_h 55) \rangle$

Les lignes 1 à 8 de notre algorithme calculent l'ensemble F_1 des items fréquents en vérifiant que le support minimum *minSupp* soit respecté. F_1 sera composé de tous les items I qui apparaissent dans au moins deux des séquences associées aux zones, et des items $\theta \cdot I$ qui apparaissent dans le voisinage d'au moins deux zones.

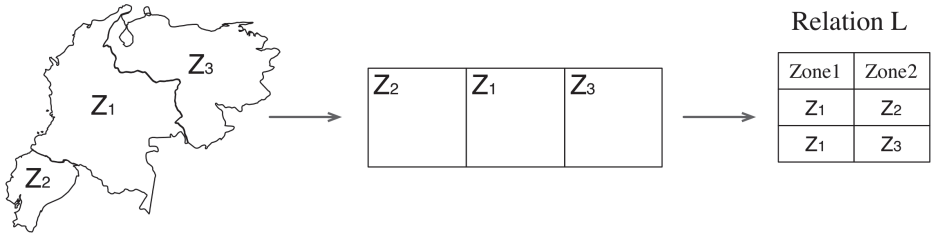


FIGURE 3.11 – Relation de voisinage L construite pour les zones représentées dans la Figure 1.2

TABLE 3.4 – Multi-ensembles de séquences et séquences voisines

Séquences	**Séquences voisines**
$s_1 = \langle (T_b H_m \mathbf{P_m})(T_m H_m P_b)(T_b H_m P_m) \rangle$	$s_2 = \langle (T_m H_m \mathbf{P_m})(T_b H_m P_b)(T_b H_b P_m 55) \rangle$
	$s_3 = \langle (T_b H_m P_h 75)(T_m H_h P_b)(T_b H_h P_h 55) \rangle$
$s_2 = \langle (T_m H_m \mathbf{P_m})(T_b H_m P_b)(T_b H_b P_m 55) \rangle$	$s_1 = \langle (T_b H_m \mathbf{P_m})(T_m H_m P_b)(T_b H_m P_m) \rangle$
$s_3 = \langle (T_b H_m P_h 75)(T_m H_h P_b)(T_b H_h P_h 55) \rangle$	$s_1 = \langle (T_b H_m \mathbf{P_m})(T_m H_m P_b)(T_b H_m P_m) \rangle$

Dans notre exemple, l'item P_m apparaît deux fois dans les séquences s_1 et s_2 et l'item $\theta \cdot P_m$ apparaît trois fois : dans le voisinage de s_1 (i.e., s_2), de s_2 (i.e., s_1) et de s_3 (i.e., s_1). Finalement, F_1 est composé de :

$$\begin{aligned} F_1 = \{ & T_b : 3, P_m : 2, P_b : 3, H_m : 3, T_m : 3, 55 : 2, \\ & \theta \cdot T_b : 3, \theta \cdot P_m : 3, \theta \cdot P_b : 3, \theta \cdot H_m : 3, \theta \cdot T_m : 3, \} \end{aligned}$$

Le nombre associé à ces items est leur support dans les séquences du Tableau 3.4.

Nous pouvons constater que l'item $\theta \cdot 55$ n'apparait pas dans l'ensemble F_1 même s'il apparait deux fois dans des séquences associées aux séquences voisines. En effet, l'item $\theta \cdot 55$ apparait dans deux séquences voisines de s_1 (i.e., s_2 et s_3), et donc dans un seul voisinage (celui de s_1), ce qui explique que son support soit de $1/3$.

À partir de l'ensemble F_1, nous construisons l'ensemble F_2 des motifs spatio-séquentiels préfixés par des motifs spatio-séquentiels trouvés à l'étape précédente. De façon générale, nous procédons comme suit : si deux motifs ont le même préfixe (ligne 13), nous générons un motif candidat γ comportant $k+1$ items qui est obtenu par "union" de deux motifs spatio-séquentiels fréquents comportant k items (ligne 14) selon les cas présentés dans le Tableau 3.2 .

Par exemple, voici quelques motifs candidats composés de 2 items, générés à partir de l'ensemble F_1 : $(T_b)(T_b)$, $(T_b P_m)$, $(T_b P_m)$, ..., $(T_b \cdot T_b)$, $(T_b)(\theta \cdot T_b)$, ... Dans cet ensemble de

candidats, nous ne gardons que ceux pour lesquels leurs sous-séquences sont fréquentes (ligne 15).

Ensuite, la ligne 16 de l'Algorithme 2 vérifie que chaque motif candidat retenu dans l'étape précédente est fréquent. Par exemple, le motif $\langle(T_m)(P_h)\rangle$ est fréquent parce qu'il apparait deux fois dans les séquences s_1 et s_2 du Tableau 3.4.

Finalement, nous obtenons l'ensemble des motifs spatio-séquentiels fréquents composés de 2 items :

$$\begin{aligned} F_2 = \{&(T_b)(T_b):3,(T_b\cdot T_b):3,(T_b)(\theta\cdot T_b):2,(T_bP_m):2,\\ &(P_m\cdot P_m):3,(H_m)(H_m):2,(H_m\cdot H_m):3,(H_m)(\theta\cdot H_m):3,\ldots\} \end{aligned}$$

En suivant toutes ces étapes de façon *itérative*, l'algorithme continue jusqu'à ce qu'il n'y ait plus aucun candidat qui puisse être généré. Il retourne ensuite l'union des ensembles F_k de motifs spatio-séquentiels.

SPADE est considéré comme un algorithme *par niveau*. La complexité de calcul de l'algorithme *BFS-S2PMiner* est, dans le pire des cas, $O(TN(2M)^N)$ où T est le nombre de transactions sur la base de données ($|dom(D_S)|$), N est le nombre d'items différents dans la base de données et M est le nombre maximal de zones voisines.

Approche de parcours en profondeur

De nombreux algorithmes utilisent une stratégie de parcours en profondeur basée sur des projections successives de la base de données tels que *PSP* [Masseglia *et al.*, 1998], *FP-Growth* [Han *et al.*, 2000] et *PrefixSpan* [Mortazavi-Asl *et al.*, 2000]. Dans cette section, nous proposons un algorithme appelé *DFS-S2PMiner* utilisant ce type d'approche pour extraire des motifs spatio-séquentiels. Plus précisément, cet algorithme est basé sur la stratégie *pattern-growth* proposée dans [Han *et al.*, 2000] et [Mortazavi-Asl *et al.*, 2000].

Le principe de ce type d'approche est d'extraire les motifs fréquents sans passer par une phase de génération des candidats. En effet, cette phase des algorithmes par niveau peut être particulièrement coûteuse en raison du nombre important de motifs candidats pouvant être générés. De plus, elle implique de parcourir la base de données de manière répétée et de vérifier le support d'un grand nombre de motifs. Les stratégies de type *pattern-growth* évitent cette phase en effectuant une approche "diviser pour régner". Cette approche fait récursivement des projections de la base de données, les associe à un fragment de motif fréquent et "fouille" chacune des bases projetées séparément. Les motifs fréquents sont ainsi étendus progressivement suivant un parcours en profondeur de l'espace de recherche.

L'Algorithme 3 décrit l'algorithme récursif *DFS-S2PMiner*. Premièrement, l'algorithme commence par construire l'ensemble F_1 des items fréquents I et $\theta\cdot I$ de la projection $seqBD|_\alpha$ (ligne 1 de l'Algorithme 3). Ces items vont constituer les extensions possibles de

Algorithme 3: DFS-S2PMiner

– Main routine

ENTRÉ: Une base de séquences $seqBD$, une relation de voisinage L
et un support minimum $minSupp$
SORTIE: L'ensemble de motifs spatio-séquentiels F
$\alpha \leftarrow \{\}$
Prefix-growthST$(\alpha, minSupp, seqDB|_\alpha, F)$

– Prefix-growthST$(\alpha, minSupp, seqDB|_\alpha, F)$

ENTRÉ: le motif spatio-séquentiel α, le support minimum $minSupp$, la α-projection $seqDB|_\alpha$
de la base de séquences $seqDB|$ et l'ensemble de motifs spatio-séquentiels F ;

1. $F_1 \leftarrow$ {l'ensemble d'items fréquents I et $\theta \cdot I$ de la projection $seqDB|_\alpha$ avec $I \in \bigcup_{i \in [1..p]} dom(D_{A_i})$}
2. **pour tout** $X \in F_1$ **faire**
3. $\quad \beta \leftarrow \alpha X$
4. $\quad \delta \leftarrow \alpha(X)$
5. $\quad$ **si** $supp(\delta) \geq minSupp$ **alors**
6. $\quad\quad F \leftarrow F \cup \delta$;
7. $\quad\quad$ Prefix-growthST$(\delta, minSupp, seqDB|_\delta, F)$
8. $\quad$ **finsi**
9. $\quad$ **si** $supp(\beta) \geq minSupp$ **alors**
10. $\quad\quad F \leftarrow F \cup \beta$;
11. $\quad\quad$ Prefix-growthST$(\beta, minSupp, seqDB|_\beta, F)$
12. $\quad$ **finsi**
13. **fin pour**

la séquence α. Notons que lors du premier appel récursif, nous avons $\alpha = \{\}$. Dans ce cas, la projection $seqBD|_\alpha$ correspond à la base de séquences $seqBD$ de départ.

Ensuite, pour chaque item $X \in F_1$, nous étendons le motif spatio-séquentiel α avec X (lignes 3 et 4). Deux types d'extension sont possibles : 1) ajouter X au dernier itemset spatial de la séquence α (ligne 3) ou 2) insérer X après (i.e., au temps suivant) le dernier itemset spatial de α (ligne 4). Une fois l'extension du motif effectuée, nous vérifions le support des séquences spatiales résultantes (lignes 5 et 9) et nous stockons les séquences spatiales fréquentes dans l'ensemble des solutions F (lignes 6 et 10).

Pour chaque motif séquentiel, l'algorithme effectue ensuite une projection de la base de données à partir de $seqBD|_\alpha$ et étend de manière récursive le motif en appelant - une fois de plus - l'algorithme (lignes 7 et 11). L'algorithme s'arrête lorsqu'il n'y a plus de projections qui puissent être générées.

Exemple Nous illustrons cet algorithme en utilisant le même exemple que dans la section précédente (Tableau 3.3, Figure 3.11 et $minSupp = 2/3$).

Iteration 1 ($\alpha = \{\}$)

Dans un premier temps, l'Algorithme 3 commence par extraire les items fréquents et les items spatiaux fréquents de la base de séquences $seqBD$ (ligne 1), soit :

$$F_1 = \{H_m:3, T_m:3, P_m:2, P_b:3, T_b:3, 55:2, \theta\cdot T_m:3,$$
$$\theta\cdot H_m:3, \theta\cdot P_m:3, \theta\cdot P_b:3, \theta\cdot T_b:3, \theta\cdot 55:3\}$$

Ensuite, la séquence α est étendue (lignes 3 et 4). Dans les lignes 5, 6 et 9, 10 nous comparons et gardons les solutions fréquentes.

Pour chaque item fréquent I et $\theta\cdot$I, l'algorithme calcule la projection de la base de données par rapport à ces items (aucune extension n'est faite ici car α est vide). Par exemple, pour l'item fréquent H_m, nous obtenons la projection (voir Tableau 3.5). Chacune de ces bases projetées est ensuite utilisée dans un appel récursif permettant de rechercher ses super-séquences fréquentes (lignes 7 et 11).

TABLE 3.5 – Base projetée pour $\langle(H_m)\rangle$

Séquences	**Séquences voisines**
$s_1 = \langle(\_\mathbf{P_m})(T_m\mathbf{H_m}P_b)(T_b\mathbf{H_m}\mathbf{P_m}55)\rangle$	$s_2 = \langle(\_P_m)(T_bH_mP_b)(T_bH_bP_m)\rangle$
	$s_3 = \langle(\_P_h75)(T_mH_hP_b)(T_bH_hP_h55)\rangle$
$s_2 = \langle(\_\mathbf{P_m})(T_b\mathbf{H_m}P_b)(T_bH_b\mathbf{P_m})\rangle$	$s_1 = \langle(\_P_m)(T_mH_mP_b)(T_bH_mP_m55)\rangle$
$s_3 = \langle(\_P_h75)(T_mH_hP_b)(T_bH_hP_h55)\rangle$	$s_1 = \langle(\_P_m)(T_mH_mP_b)(T_bH_mP_m55)\rangle$

Iteration 2 ($\alpha = \langle(H_m)\rangle$)

Le premier appel récursif va construire les super-séquences ayant pour préfixe $\langle(H_m)\rangle$ à partir de la base projetée représentée par le Tableau 3.5. Plus précisément, l'algorithme va d'abord rechercher les items fréquents dans cette base projetée (lignes 1), puis les étendre $\langle(H_m)\rangle$ (lignes 3 et 4). Les items fréquents obtenus pour $seqBD|_{\langle(H_m)\rangle}$ sont : $\{P_m:2, T_m:2, H_m:2, P_b:3, T_b:3, 55:2, \theta\cdot P_m:3, \theta\cdot T_b:3, \theta\cdot H_m:3, \theta\cdot P_b:3, \theta\cdot T_m:3, \theta\cdot 55:3\}$

Le premier item fréquent trouvé est $\langle P_m\rangle$: 2. On obtient donc les motifs spatio-séquentiels $\langle(H_mP_m)\rangle$: 1 (ligne 3) et $\langle(H_m)(P_m)\rangle$: 2 (ligne 4). Seul $\langle(H_m)(P_m)\rangle$: 2 est fréquent avec un *minSupp* de 2/3 (ligne 9). L'algorithme stocke ce motif (ligne 10) l'utilise pour faire une nouvelle projection (cf., Tableau 3.6) et rechercher récursivement toutes les super-séquences fréquentes ayant pour préfixe $\langle(H_m)(P_m)\rangle$.

TABLE 3.6 – Base projetée pour $\langle(H_m)(P_m)\rangle$

Zones	**Séquences**	**Zones voisines**	**Séquences voisines**
Z_1	$s_1 = \langle(T_mH_mP_b)(T_bH_mP_m55)\rangle$	Z_2	$s_2 = \langle(T_b\mathbf{H_m}P_b)(T_b\mathbf{H_m}P_m)\rangle$
		Z_3	$s_3 = \emptyset$
Z_2	$s_2 = \langle(T_bH_mP_b)(T_bH_bP_m)\rangle$	Z_1	$s_1 = \langle(T_m\mathbf{H_m}P_b)(T_b\mathbf{H_m}P_m55)\rangle$
Z_3	$s_3 = \emptyset$	Z_1	$s_1 = \langle(T_m\mathbf{H_m}P_b)(T_b\mathbf{H_m}P_m55)\rangle$

Iteration 3 ($\alpha = \langle(H_mP_m)\rangle$)

Les items fréquents obtenus à partir de la projection $seqBD|_{\langle(H_m P_m)\rangle}$ sont $\{P_m : 2, H_m : 2, P_b : 2, T_b : 2, \theta \cdot P_m : 3, \theta \cdot T_b : 3, \theta \cdot H_m : 3, \theta \cdot P_b : 3, \theta \cdot 55 : 3\}$.

Ensuite, on étend la séquence α (lignes 3 et 4). Par exemple, l'item spatial $\theta \cdot H_m : 3$ est un des items fréquents. Dans ce cas, l'algorithme construit la séquence spatiale $\langle(H_m P_m)(\theta \cdot H_m)\rangle : 1$. Cette séquence est fréquente car elle apparait dans toutes les séquences du Tableau 3.5).

Finalement, une projection et un appel récursif (lignes 7 et 11) sont réalisés. Quand tous les items fréquents sont projetés, l'algorithme va parcourir une autre "branche" de l'espace de recherche, i.e., les motifs commençant par $\langle(T_m)\rangle$. Par exemple, l'un des items fréquents dans les séquences voisines (cf., Tableau 3.6) est $\langle P_b : 3\rangle$ donc, le motif $\langle(H_m P_m)(\theta \cdot H_m P_b)\rangle$ sera extraite dans le itération 4.

L'algorithme procède donc globalement de la même manière que les items soient spatiaux ou non. La principale différence se situe dans la manière de compter le support. En effet, le support d'un item spatial représente le nombre de zones ayant au moins une fois dans leur voisinage l'item en question (c'est pour cette raison que nous avons $\theta \cdot 55 : 1$ dans le Tableau 3.5). Notons que lorsque l'algorithme étend un motif du type $\langle(I_{ST_1})(I_{ST_2})...(I_{ST_K} \cdot X)\rangle$ avec un item fréquent $\theta \cdot Y$, l'opérateur de groupement n-aire est utilisé afin de représenter la séquence sous la forme $\langle(I_{ST_1})(I_{ST_2})...(I_{ST_K} \cdot [X;Y])\rangle$.

La complexité de calcul de l'algorithme *DFS-S2PMiner*, au pire du cas, est $O((2NM)^L)$ où N est le nombre d'items différents dans la base de données, M est le nombre maximum des zones voisines et L est la longueur maximale de toutes les transactions. La constante 2 est introduite puisque chaque élément peut être ajouté en opération, soit par itemset ou par extension de la séquence.

3.4 Discussion

Dans ce chapitre, nous avons présenté deux méthodes d'extraction de motifs spatio-temporels. Dans la première méthode, nous avons appliqué un algorithme classique d'extraction de motifs séquentiels en prenant en compte les informations spatiales (e.g. la distance entre entités spatiales, l'appartenance, etc.). Nous avons souligné les problèmes posés par les différentes spatialisations et leur influence sur le nombre de motifs extraits.

Dans la deuxième approche, nous avons donné un cadre théorique pour l'extraction de motifs spatio-séquentiels, des séquences d' itemsets spatiaux, représentant l'évolution dans le temps de zones et de leur environnement. Pour extraire ce type de motifs à partir d'une base de données spatio-temporelles, nous avons proposé deux algorithmes d'extraction selon les deux paradigmes les plus classiques en extraction de connaissances. Le premier effectue un parcours par niveau de l'espace de recherche en s'appuyant sur une stratégie générer-élaguer-tester. Le second effectue un parcours en profondeur en s'appuyant sur des projections successives de la base de données. Ce type de motifs génère un important espace de recherche. Pour faire face à ce problème, nous avons introduit

également deux mesures d'intérêt adaptées aux aspects spatio-temporels de ces motifs. La première est une extension du support tandis que la deuxième est une extension de la mesure appelée *indice de participation en cascade* proposé par Mohan *et al.* [2012].

Maintenant, le problème posé est de savoir si tous les motifs obtenus en utilisant les deux méthodes sont intéressants pour un expert. Dans la suite, nous allons nous focaliser sur cette question en proposant deux mesures de qualité.

Chapitre 4

Mesures de qualité

Préambule

Après l'étape de fouille de données, les motifs extraits sont présentés aux utilisateurs finaux. Très souvent, le nombre de motifs extraits est trop important pour permettre une analyse manuelle et parfois même supérieur au nombre de données en entrée du processus. Il est donc crucial de ne conserver que les motifs les plus pertinents. La pertinence des motifs dépend des besoins des utilisateurs qui peuvent s'intéresser aux motifs fréquents, rares, inattendus, etc. Dans ce chapitre, nous allons proposer deux mesures permettant de filtrer les motifs en utilisant la notion de "contradiction".

4.1 Introduction

Un problème récurrent en fouille de données, est celui du grand nombre de motifs obtenus. Ils peuvent être parfois beaucoup plus nombreux que le nombre de séquences en entrée du processus. Or, certains de ces motifs peuvent être peu intéressants pour les experts. Ces deux problèmes sont souvent associés à l'utilisation du support comme seule mesure d'élagage. D'un côté, le support ne permet pas de détecter efficacement les tendances dans les séquences si le motif est soumis au bruit. D'un autre côté, pour de longues séquences, les algorithmes de fouille de données peuvent générer un grand nombre de motifs redondants et parfois triviaux dans lesquels, les motifs les plus pertinents sont souvent cachés parmi des motifs peu intéressants [Kum *et al.*, 2005].

Par ailleurs, comme discuté dans Baesens *et al.* [2000], le choix de la valeur optimale du support minimum dans le processus de fouille de données est une tâche très difficile

pour l'utilisateur car elle peut avoir un impact direct sur le nombre de motifs à extraire. Deux cas sont possibles : (1) pour une valeur de support haute (proche de 1), des motifs très généraux seront extraits et les motifs rares (peu fréquents), mais pouvant être intéressants, ne sont pas générés ; et (2) pour un support bas, les motifs fréquents et non fréquents sont générés mais l'ensemble peut être difficile à interpréter du fait du nombre important de motifs dont certains peuvent ne pas être pertinents pour les experts. En conséquence, choisir un support permettant d'extraire des motifs intéressants mais en petit nombre est une tâche très complexe. De plus, si les attributs des données sont fortement corrélés, le nombre de motifs est élevé [Baesens *et al.*, 2000].

Cette très grande quantité de motifs peut contenir du bruit, e.g. des motifs qui ont une très petite taille (dans notre contexte, une séquence de taille 1, i.e., contenant un seul itemset, n'exprime pas la notion d'évolution temporelle), ou ceux qui ne contiennent pas un évènement spécifique (dans notre exemple, des motifs ne contenant pas l'évènement climatique "orage"). Ces contraintes peuvent être facilement ajoutées dans l'étape de fouille de données. En effet, lorsqu'on conçoit l'algorithme d'extraction de motifs, en plus de la mesure d'élagage, d'autres contraintes peuvent être incluses, permettant ainsi de réduire le nombre de motifs et d'augmenter leur pertinence (c.f., Section 3.3). Néanmoins, filtrer les motifs les plus pertinents, reste une tâche difficile dont le post-traitement semble une alternative pertinente [Geng et Hamilton, 2006].

Bruha et Famili [2000] définissent le post-traitement comme l'*ensemble des étapes incluant des routines d'élagage, de filtrage de motifs ou d'intégration de connaissances, permettant de retirer la connaissance imprécise extraite à la suite de l'étape de fouille de données*. Le post-traitement permet aussi de simplifier la connaissance extraite, de l'évaluer et de vérifier à quel point les motifs extraits contribuent à la solution du problème initialement identifié. D'auprès ces auteurs, les différentes procédures qui pourront être appliquées dans l'étape de post-traitement peuvent être regroupées en quatre catégories : (1) filtrage des connaissances ; (2) évaluation ; (3) intégration de connaissances ; et (4) interprétation et éclaircissement.

Conjointement aux procédures citées auparavant, il existe plusieurs stratégies à mettre en œuvre au cours du post-traitement. Ces stratégies sont organisées en deux groupes : (1) la première est axée sur les données (data-driven) et utilise des mesures objectives ; et (2) la deuxième utilise des mesures subjectives et est axée sur l'utilisateur (user-driven) [Freitas, 1998; Silberschatz et Tuzhilin, 1996]. Une mesure objective évalue la pertinence d'un motif en termes de structure et selon les données utilisées pour sa construction. Dans le contexte des règles d'association, le support d'une règle est une mesure objective [Agrawal *et al.*, 1993]. Contrairement, une mesure subjective n'évalue pas seulement la structure d'un motif et les données utilisées en entrée, elle évalue aussi les possibles contraintes associées aux choix des utilisateurs. Par exemple, avec une mesure objective, nous vérifions si un motif apparait au moins dans 50% des quartiers d'une ville (si le support minimum est fixé à 0,5), avec une mesure subjective nous pouvons extraire ceux qui contiennent un

évènement intéressant pour les utilisateurs, dans notre cas d'étude, les motifs contenant l'évènement *présence d'orages* par exemple.

Dans ce chapitre, nous nous focalisons sur le filtrage des motifs à l'aide d'une mesure axée sur les données. Pour cela, nous allons étendre une mesure connue dans le domaine des règles d'association et appelée *la moindre contradiction*. Nous l'appliquerons, dans un premier temps, aux motifs séquentiels et dans un deuxième temps, aux motifs spatio-séquentiels.

4.2 État de l'art

De nombreuses mesures existent pour filtrer les motifs les plus pertinents. Zaki et Hsiao [1999] présentent des premiers travaux dans lesquels un processus de post-traitement a été utilisé. Dans cet objectif, les auteurs ont généré les itemsets fermés fréquents à partir de l'ensemble des itemsets fréquents maximaux (par rapport à l'inclusion ensembliste) [1]. Pour cela, les auteurs ont cherché tous les sous-ensembles d'itemsets maximaux, en éliminant un itemset si son support est égal à l'un de ses sous-ensembles. À titre de rappel, les motifs fréquents fermés et maximaux sont des sous-ensembles de motifs fréquents mais les itemsets fréquents maximaux sont une représentation plus compacte correspondant à un sous-ensemble de motifs fréquents fermés.

Dans le contexte des règles d'association (sous la forme antécédent $\rightarrow$ conséquent), de nombreuses mesures ont été proposées pour évaluer la pertinence des règles extraites et réduire l'ensemble des solutions qu'un expert humain doit analyser [Bogorny *et al.*, 2008]. Une étude comparative de plusieurs mesures de qualité est présentée par Tan *et al.* [2002].

Les mesures classiques des règles d'association permettent d'évaluer l'écart d'indépendance entre l'antécédent et le conséquent. La confiance [Agrawal et Srikant, 1995], le nombre de contre-exemples associés à des règles [Azé, 2003], l'étonnement statistique [Lerman et Azé, 2007] sont quelques exemples de ces mesures. Une autre approche de post-traitement a été proposée par Hussain *et al.* [2000] dans laquelle, les auteurs identifient, à partir d'un ensemble de règles, un sous-ensemble appelé *règles d'exception*. Une règle quelconque est une règle d'exception, dénotée par $AB \rightarrow \neg X$, décrivant les évènements X et $\neg X$ étant donné AB. Par ailleurs, Zhao *et al.* [2008] définissent des règles d'association particulières appelées *motifs combinés*. Un motif combiné est une règle composée de multiples itemsets hétérogènes provenant de différents sources de données. Par exemple, un motif combiné peut être composé des données transactionnelles associées au marketing, à la démographie, etc. Pour filtrer ce type de motifs, les auteurs ont proposé une extension de la mesure *lift* [2] comme mesure d'élagage. Une fois les motifs combinés extraits, une deuxième étape a été proposée afin d'extraire des *paires de règles combinées*.

1. Un itemset est maximal si aucun de ses sur-ensembles est fréquent.

2. Le *lift* mesure combien de fois deux évènements X et Y se produisent souvent ensemble s'ils sont statistiquement indépendants.

Une paire de règles combinées est composée de deux règles contrastées et peuvent refléter des comportements du type : *des clients ayant les mêmes caractéristiques et différentes politiques/campagnes, peuvent entraîner deux comportements différents*. Finalement, une étape de "clustering" de motifs combinés a été proposée. Pour ces deux dernières étapes, les auteurs ont proposé une mesure de dissimilarité basée sur le *lift*.

Concernant les motifs séquentiels, Saneifar *et al.* [2008] proposent une mesure de similarité permettant de comparer deux motifs séquentiels au niveau des items et des itemsets. Les auteurs ont proposé également une extension de la méthode de clustering *k-means* pour les motifs séquentiels. De même, Prinke *et al.* [2006] ont proposé une étape de post-traitement en adaptant une mesure initialement proposée pour les règles d'association appelée *mesure d'amélioration* et qui ne peut pas être applicable aux séquences d'itemsets. La mesure proposée vise à réduire le nombre de motifs séquentiels extraits en capturant la différence minimale entre le support du motif et le support d'une super-séquence contenant le motif. Une valeur élevée de la mesure d'amélioration signifie que, le fait d'ajouter des éléments au motif a provoqué une diminution significative de son support.

Pour les données spatio-temporelles, plusieurs mesures ont été proposées. Par exemple, Sengstock *et al.* [2012] font une étude des mesures d'intérêt pour les co-locations. Ils proposent deux mesures, l'une basée sur l'entropie et l'autre basée sur la divergence. La première exploite la notion de répartition d'un motif dans l'espace tandis que la deuxième est basée sur une mesure de probabilité conditionnelle. Par ailleurs, Yoo et Bow [2011] ont proposé une étape de post-traitement qui génère les *top-k* co-locations fermées [3], i.e., les k co-locations les plus représentatives des co-locations fréquentes. Ainsi, à partir des co-locations fermées, on peut déduire le support de n'importe quelle co-location fréquente sans recourir au parcours de la base de transactions.

Bien que ces mesures appliquées aux données spatio-temporelles aient reçu beaucoup d'attention de la communauté de fouille de données [Jalali-Heravi et Zaïane, 2010], à notre connaissance, aucune mesure n'a été proposée pour évaluer la pertinence d'un motif extrait d'une base de données spatio-temporelles d'une manière facilement compréhensible par les experts, tout en prenant en compte les aspects temporel et/ou spatial des données.

Dans la suite, nous décrivons deux mesures : la moindre contradiction temporelle pour les motifs séquentiels et la moindre contradiction spatio-temporelle pour les motifs spatio-séquentiels.

4.3 La moindre contradiction temporelle

Comme expliqué dans l'état de l'art, les mesures utilisées dans le cas des règles d'association se basent pour la plupart sur l'indépendance entre l'antécédent et le conséquent de la règle.

3. Une co-location fermée est un ensemble maximal de type d'évènements partagés pour un ensemble de transactions d'une base de données spatio-temporelles.

Ces mesures ne peuvent pas s'appliquer aux motifs séquentiels car elles n'intègrent pas la notion d'ordre entre les itemsets. À l'exception du support, les mesures de qualité existantes ne peuvent donc pas être utilisées directement.

Nous avons ainsi choisi de nous focaliser sur la recherche de séquences fréquentes qui sont peu contredites par les données. Pour cela, nous proposons d'étendre la mesure de *moindre contradiction* (MC), définie par Azé [2003] pour les règles d'association, au contexte des séquences d'itemsets. Rappelons que cette mesure, dans le cadre d'une règle d'association $A \rightarrow B$ où A et B sont deux ensembles d'items disjoints, est définie par :

$$MC(A \rightarrow B) = \frac{supp(AB) - supp(A\overline{B})}{supp(B)} \tag{4.1}$$

où $A\overline{B}$ est l'itemset tel que A est présent et B absent.

Bien que cette mesure ne soit pas la seule à présenter de telles propriétés, nous avons choisi d'étendre la moindre contradiction aux séquences d'itemsets pour deux raisons principales. Premièrement, cette mesure est simple à comprendre et à mettre en œuvre. Elle est donc relativement simple à appréhender par les experts. Deuxièmement, des travaux précédents ont montré que cette mesure permet d'extraire des pépites de connaissances [Azé, 2003] et qu'elle résiste au bruit [Azé *et al.*, 2007], *etc.* Enfin, d'autres mesures pourront également être étendues aux séquences temporelles comme le *lift* dont la définition est "proche" de celle de la moindre contradiction.

4.3.1 Définitions

Dans cette sous-section, nous allons donner quelques définitions concernant la mesure que nous proposons ainsi que le détail de son calcul.

Définition 4.1 ***La moindre contradiction temporelle*** *Soit une séquence fréquente (motif séquentiel)* s *appartenant à la base de motifs* mBD*, la moindre contradiction temporelle de* s*, notée* $MCT(s, mBD)$ *est définie par :*

$$MCT(s, mBD) = \frac{supp(s) - \sum_{sc \in S_{contr}} supp(sc)}{\sum_{sa \in S_{all}} supp(sa)} \tag{4.2}$$

où $\begin{cases} S_{contr} & \textit{l'ensemble des séquences de } mBD \textit{ incluant tous les itemsets de la séquence } s \textit{ mais dans un ordre différent} \\ S_{all} & \textit{l'ensemble des séquences de } mBD \textit{ incluant tous les items qui sont apparus dans la séquence } s \end{cases}$

Cette extension de la moindre contradiction permet de conserver l'esprit initial de la mesure qui vise à évaluer le nombre de fois où une règle est vérifiée *vs* le nombre de fois où elle est invalidée. Une règle qui est plus fréquemment vérifiée qu'invalidée est a priori intéressante. Comme pour la version "classique", cette mesure est normalisée. Ici, la normalisation est effectuée en utilisant le support total des séquences contenant les mêmes items.

Exemple :
Soit la base de motifs mBD contenant les séquences suivantes et leur support :

$$\begin{cases} s_1 = \langle (AB)(BC) \rangle & supp(s_1) = 0,25 \\ s_2 = \langle (BC)(AB) \rangle & supp(s_2) = 0,10 \\ s_3 = \langle (AB)(CE) \rangle & supp(s_3) = 0,12 \\ s_4 = \langle (AB) \rangle & supp(s_4) = 0,13 \\ s_5 = \langle (EA)(BC) \rangle & supp(s_5) = 0,20 \end{cases}$$

Alors,

$$MCTS(s_1, mBD) = \frac{supp(s_1) - \sum\limits_{sc \in S_{contr}} supp(sc)}{\sum\limits_{sa \in S_{all}} supp(sa)} = \frac{0,25 - 0,10}{0,67} = 0,224$$

$$avec \begin{cases} supp(s_1) = 0,25 \\ S_{contr} = \{s_2\} \\ S_{all} = \{s_1, s_2, s_3, s_5\} \end{cases}$$

On retrouve (BC) et (AB) dans S_2 (qui a les mêmes itemsets que la séquence S_1 mais dans un ordre différent) et on retrouve les items A, B et C dans S_1, S_2, S_3 et S_5, mais pas dans S_4 qui ne contient que les items A et B.

4.3.2 Algorithme

L'Algorithme 4 décrit les étapes que nous proposons pour le calcul de *la moindre contradiction temporelle*. Cet algorithme se déroule en deux étapes. Tout d'abord, l'algorithme recherche les séquences contenant des itemsets communs entre la séquence traitée et la séquence candidate sans considérer l'ordre d'apparition (lignes 7 à 16 de l'Algorithme 4). Ensuite, il recherche tous les items appartenant à la séquence traitée dans la séquence candidate (lignes 18 à 24 de l'Algorithme 4).

Cet algorithme a une complexité d'au plus $O(n^2)$ où n représente le nombre de n-uplets de la base de motifs séquentiels.

Algorithme 4: Calcul de la moindre contradiction temporelle

ENTRÉ: mBD : Base de motifs séquentiels et leurs supports
SORTIE: la moindre contradiction temporelle de chaque séquence $s \in mBD$

```
1.  vecteur MCT ;
2.  pour tout (séquence s1 ∈ mBD) faire
3.      Scontr ← 0 ;
4.      Sall ← 0 ;
5.      pour tout (séquence s2 ∈ mBD) faire
6.          boolean all_in ← vrai ;
7.          tant que ((all_in) et (∀ itemsets IS ∈ s1)) faire
8.              si (IS ⊄ s2) alors
9.                  all_in ← faux ;
10.             sinon
11.                 prochain IS ∈ s1 ;
12.             finsi
13.         fin tant que
14.         si (all_in) alors
15.             Scontr ← Scontr + supp(s2) ;
16.         finsi
17.         boolean all_in ← vrai ;
18.         tant que ((all_in) et (∀ item I ∈ s1)) faire
19.             si (I ⊄ s2) alors
20.                 all_in ← faux ;
21.             sinon
22.                 prochain I ∈ s1 ;
23.             finsi
24.         fin tant que
25.         si (all_in) alors
26.             Sall ← Sall + supp(s2);
27.         finsi
28.         MCT(s1) ← (supp(s1) − Scontr) / Sall ;
29.     fin pour
30.     retourner MCT;
31. fin pour
```

4.4 La moindre contradiction spatio-temporelle

Nous avons également étendu la mesure de qualité définie par Azé [2003] pour les motifs spatio-séquentiels. Cette mesure, que nous avons appelée *moindre contradiction spatio-temporelle*, conserve l'esprit de la mesure mentionnée précédemment et a pour rôle de filtrer les motifs spatio-séquentiels les moins contradictoires par rapport à la base de motifs.

Dans ce contexte, nous avons pris en compte trois contraintes : (1) la dimension temporelle des itemsets spatiaux ; (2) la dimension spatiale d'itemsets spatiaux (i.e., l'ensemble

des évènements produits dans une zone et son environnement proche à un moment donné) ; et (3) la symétrie des itemsets spatiaux, i.e., ceux de la forme $(A \cdot B)$ et $(B \cdot A)$.

Par rapport aux deux premières contraintes, nous allons prendre en compte la séquentialité des itemsets spatiaux et la relation de voisinage au moment de calculer la valeur de la moindre contradiction spatio-temporelle associée à un motif.

Concernant la troisième contrainte, nous voudrions également, filtrer les itemsets spatiaux symétriques. Pour éclaircir la notion de symétrie d'itemsets spatiaux, prenons l'exemple de la Figure 4.1. L'itemset spatial $(T_b \cdot P_b)$ peut être interprété par : dans la zone Z_1, l'évènement *température basse* apparaît en même temps que l'évènement *précipitation basse* dans une zone voisine à Z_1. De la même façon, l'itemset spatial $(P_b \cdot T_b)$ représente l'apparition de l'évènement *précipitation basse* dans Z_2 et *température basse* dans leur entourage. Même si ces deux itemsets spatiaux peuvent être interprétés de deux façons différentes, sémantiquement ils représentent le même phénomène, i.e., *température basse* et *précipitation basse* sont deux évènements qui apparaissent fréquemment dans deux zones voisines à un temps donné. Grâce à la moindre contradiction spatio-temporelle, nous allons pondérer des itemsets spatiaux non symétriques.

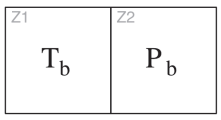


FIGURE 4.1 – Symétrie des itemsets spatiaux

Notre mesure permettra donc de mettre en évidence des évènements apparus dans une zone à différents moments et dans leur entourage proche. Elle permettra également de filtrer les motifs contenant des itemsets symétriques diminuant ainsi la redondance.

Dans les sections suivantes, nous allons présenter les définitions associées à la moindre contradiction spatio-temporelle, l'algorithme permettant de calculer cette mesure ainsi que quelques illustrations.

4.4.1 Définitions

Définition 4.2 ***Similarité entre séquences spatiales***

Deux itemsets spatiaux I_{ST_1} *et* I_{ST_2} *sont dits "similaires" et représentés par* $similaire(I_{ST_1}, I_{ST_2})$ *si et seulement si, ils contiennent les mêmes items. Plus formellement,*

$$\begin{cases} similaire(I_{ST_1}, I_{ST_2}) = vrai \textit{ si } \forall x (x \in I_{ST_1} \leftrightarrow x \in I_{ST_2}) \\ similaire(I_{ST_1}, I_{ST_2}) = faux \textit{ sinon} \end{cases} \tag{4.3}$$

$$\textit{où } x \in dom(D_{A_i})$$

Par exemple, les itemsets spatiaux $(AB \cdot C)(C \cdot AB)$ et (ABC) sont similaires.

De la même façon, nous allons définir la fonction $pos(I_{ST}, s) \rightarrow n$ où I_{ST} est un itemset spatial, s est une séquence spatiale et n est un nombre entier. La fonction $pos(I_{ST}, s)$ retourne la position de l'itemset I_{ST} dans la séquence spatiale s.

Par exemple, soit la séquence spatiale $s = \langle (A \cdot BC)(CD) \rangle$, donc, $pos((CD), s) = 1$.

Définition 4.3 ***La moindre contradiction spatio-temporelle***

Soit s *une séquence spatiale fréquente appartenant à la base de motifs* mBD*, la moindre contradiction spatio-temporelle de* s*, notée par* $MCST(s, mBD)$*, est définie par :*

$$MCST(s, mBD) = \frac{supp(s) + \sum\limits_{\min_{\subseteq}\{ss \in S_{sim}\}} supp(ss) - \sum\limits_{\min_{\subseteq}\{sc \in S_{contr}\}} supp(sc)}{\sum\limits_{\min_{\subseteq}\{sa \in S_{all}\}} supp(sa)} \tag{4.4}$$

où :

- S_{sim} *l'ensemble des séquences spatiales de* mBD *incluant tous les itemsets spatiaux similaires de la séquence* s *dans le même ordre*
- S_{contr} *l'ensemble des séquences spatiales de* mBD *incluant tous les itemsets spatiaux similaires de la séquence* s *mais dans un ordre différent*
- S_{all} *l'ensemble des séquences de* mBD *incluant tous les items qui sont apparus dans la séquence* s

Exemple :

Soit la base de motifs mBD contenant les séquences spatiales suivantes et leur support :

$$\begin{cases} s_1 = \langle (AB)(A \cdot BC) \rangle & supp(s_1) = 0{,}42 \\ s_2 = \langle (B \cdot B)(A \cdot C)(C) \rangle & supp(s_2) = 0{,}35 \\ s_3 = \langle (A \cdot BC)(AB) \rangle & supp(s_3) = 0{,}30 \\ s_4 = \langle (A \cdot BC)(\theta \cdot C)(AB) \rangle & supp(s_4) = 0{,}25 \\ s_5 = \langle (A)(AB)(B)(A \cdot BC) \rangle & supp(s_5) = 0{,}20 \\ s_6 = \langle (AB)(BC \cdot A)(A \cdot B) \rangle & supp(s_5) = 0{,}17 \end{cases}$$

Alors,

$$MCTS(s_1, mBD) = \frac{0{,}42 + [0{,}20 + 0{,}17] - [0{,}30 + 0{,}17]}{[0{,}35 + 0{,}30 + 0{,}20 + 0{,}17]} = 0{,}25197$$

$$avec \begin{cases} supp(s_1) = 0{,}42 \\ S_{sim} = \{s_5, s_6\} \\ S_{contr} = \{s_4, s_6\} \\ S_{all} = \{s_2, s_4, s_5, s_6\} \end{cases}$$

D'abord, nous cherchons l'ensemble des séquences spatiales comprenant tous les itemsets spatiaux de s_1 qui sont apparus dans le même ordre temporel. Cet ensemble est composé par les séquences s_5 et s_6.

Après, nous recherchons des séquences spatiales contenant tous les itemsets spatiaux de s_1, mais dans un ordre temporel différent. Les séquences s_4 et s_6 ont été maintenu dans cette ensemble. À ce stade, il est important de noter que la séquence spatiale s_3 contredit aussi s_1. Néanmoins, étant donné que $s_3 \subseteq s_4$, nous ne considérons que le support de s_4, car les séquences de données supportant s_4 et contredisant s_1, sont inclues dans celles supportant s_3. Il ne faudrait donc pas les compter deux fois.

Enfin, nous recherchons des séquences contenant tous les éléments de s_1 i.e., $(A, B, \cdot B$ et $\cdot C)$. Ainsi, nous avons trouvé tous ces objets dans des séquences s_2, s_4, s_5 et s_6. La séquence s_3 n'a été pas gardée car $s_3 \subseteq s_4$ ainsi que la séquence s_1 car $s_1 \subseteq s_6$.

Dans l'exemple, la séquence s_6 est incluse dans l'ensemble S_{sim} car $\langle(AB)(BC\cdot A)(\ldots)\rangle$ et dans l'ensemble S_{contr} car $\langle(\ldots)(BC\cdot A)(A\cdot B)\rangle$

4.4.2 Algorithme

L'Algorithme 5 décrit les trois étapes du calcul de *la moindre contradiction spatio-temporelle* : (1) tout d'abord, nous cherchons les séquences contenant des itemsets spatiaux "similaires" entre la séquence traitée et la séquence candidate en prenant en compte l'ordre d'apparition (lignes 8 à 17 de l'Algorithme 5) ; (2) ensuite, nous évaluons les séquences contenant des itemsets spatiaux "similaires" entre la séquence traitée et la séquence candidate sans considérer l'ordre d'apparition (lignes 19 à 28) ; et finalement (3) nous recherchons tous les items appartenant à la séquence traitée dans la séquence candidate (lignes 30 à 39 de l'Algorithme 5). Au cours de ces trois étapes, nous vérifions la contrainte d'inclusion des séquences spatiales et gardons les séquences ayant le support le plus petit.

Cet algorithme a une complexité d'au plus $O(\log(n) * n^2)$ où n représente le nombre de n-uplets de la base de motifs. Il faut noter que le temps d'exécution de l'Algorithme 5 augmente considérablement par rapport à son prédécesseur en raison de la longueur des séquences spatiales. En effet, les séquences spatiales contiennent un nombre d'évènements considérablement supérieur à celui des séquences temporelles car on considère les évènements apparus dans des zones voisines.

4.5 Discussion

Dans ce chapitre, nous avons présenté deux mesures de qualité permettant de filtrer les nombreux motifs obtenus dans l'étape de fouille. Pour cela, nous avons étendu la mesure appelée *la moindre contradiction*. Cette mesure a pour objectif de filtrer les motifs les plus contradictoires par rapport aux données. Nous avons donc étendu la moindre contradic-

Algorithme 5: Calcul de la moindre contradiction spatio-temporelle

ENTRÉ: mBD : Base de motifs spatio-séquentiels et leurs supports
SORTIE: *la moindre contradiction spatio-temporelle* de chaque motif $s \in mBD$

1. vecteur MCST ;
2. **pour tout** (séquence $s_1 \in mBD$) **faire**
3. $S_{sim} \leftarrow 0$;
4. $S_{cont} \leftarrow 0$;
5. $S_{all} \leftarrow 0$;
6. **pour tout** (séquence $s_2 \in mBD$) **faire**
7. boolean $all\_in \leftarrow vrai$;
8. **tant que** ((all_in) **et** ($\forall$ itemset $IS_1 \in s_1$)) **faire**
9. **tant que** ((all_in) **et** ($\forall$ itemset $IS_2 \in s_2$)) **faire**
10. **si** $(similar(IS_1, IS_2) = faux)$ **alors**
11. $all\_in \leftarrow faux$;
12. **finsi**
13. **fin tant que**
14. **fin tant que**
15. **si** $((all\_in)$ **et** $(s_2 \not\subset s_m \in MCST)$ **et** $(supp(s_2) < supp(s_m)))$ **alors**
16. $S_{sim} \leftarrow S_{sim} + supp(s_2)$;
17. **finsi**
18. $all\_in \leftarrow vrai$;
19. **tant que** ((all_in) **et** ($\forall$ itemset $IS_1 \in s_1$)) **faire**
20. **tant que** ((all_in) **et** ($\forall$ itemset $IS_2 \in s_2$)) **faire**
21. **si** $((similar(IS_1, IS_2) = faux)$ **et** $(pos(IS_1, s_1) = pos(IS_2, s_2))$ **alors**
22. $all\_in \leftarrow faux$;
23. **finsi**
24. **fin tant que**
25. **fin tant que**
26. **si** $((all\_in)$ **et** $(s_2 \not\subset s_m \in MCST)$ **et** $(supp(s_2) < supp(s_m)))$ **alors**
27. $S_{contr} \leftarrow S_{contr} + supp(s_2)$;
28. **finsi**
29. $all\_in \leftarrow vrai$;
30. **tant que** ((all_in) **et** ($\forall$ item $I \in s_1$)) **faire**
31. **si** $(I \not\subset s_2)$ **alors**
32. $all\_in \leftarrow faux$;
33. **sinon**
34. prochain $I \in s_1$;
35. **finsi**
36. **fin tant que**
37. **si** $((all\_in)$ **et** $(s_2 \not\subset s_m \in MCST)$ **et** $(supp(s_2) < supp(s_m)))$ **alors**
38. $S_{all} \leftarrow S_{all} + supp(s_2)$;
39. **finsi**
40. $MCST(s_1) \leftarrow \frac{supp(s_1) + S_{sim} - S_{contr}}{S_{all}}$;
41. **fin pour**
42. **retourner** MCST;
43. **fin pour**

tion pour filtrer les motifs spatio-séquentiels "intéressants", tout en gardant la nature de cette mesure. Dans un premier temps, nous avons pris en compte l'ordre temporel d'apparition des évènements et dans un deuxième temps, nous avons également pris en compte la dimension spatiale de ces motifs.

Dans le chapitre suivant, nous allons appliquer les propositions que nous avons décrites dans les deux chapitres précédents sur des données réelles et synthétiques. Ces expérimentations ont pour objectif de valider nos contributions d'un point de vue qualitatif et quantitatif.

Chapitre 5

Applications à des données réelles

Préambule

Dans ce chapitre, nous allons mener des expérimentations sur deux jeux de données réelles pour montrer l'intérêt des motifs proposés dans cette thèse et sur des jeux de données synthétiques pour évaluer leurs performances.

5.1 Introduction

Comme nous l'avons décrit dans le Chapitre 1, de nombreux phénomènes spatio-temporels rencontrés dans notre vie quotidienne peuvent être représentés à l'aide d'une base de données spatio-temporelles. Les méthodes présentées dans ce manuscrit ont été conçues pour étudier certains phénomènes très complexes, dans lesquels, des caractéristiques ou évènements décrivant une zone évoluent au cours du temps. Les zones étudiées peuvent être mises en relation grâce à des opérateurs spatiaux, comme le voisinage ou le groupement des zones selon des critères d'homogénéisation.

Comme souligné par Bergeret *et al.* [2007], un algorithme doit être testé sur des données réelles ainsi que sur des données synthétiques pour confirmer les résultats théoriques envisagés. Dans ce but, pour tester la performance et le passage à l'échelle de nos algorithmes, nous avons développé un générateur de données synthétiques paramétrable permettant d'obtenir des données auto-générées réalistes au sens où elles s'approchent le plus des caractéristiques de nos jeux de données réelles tout en permettant de tester des cas limites. Nous confronterons également nos méthodes à deux jeux de données réelles issus d'une étude sur la pollution des rivières et d'une étude sur les épidémies de dengue.

Dans les sections suivantes, nous allons dans un premier temps, décrire en détail les jeux de données réelles ainsi que le générateur de données synthétiques utilisé au cours de cette thèse. Ensuite, nous allons évaluer nos propositions grâce à une vaste collection de tests sur ces différents jeux de données. La Figure 5.1 illustre les étapes suivies pour les expérimentations décrites dans ce chapitre (quatre premiers blocs de la Figure 5.1).

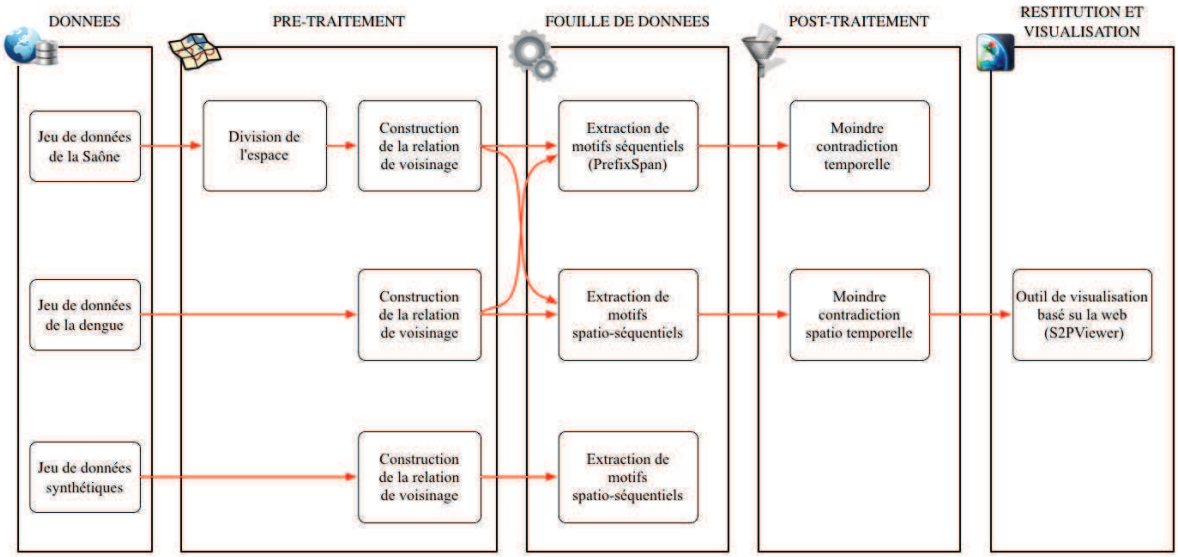


FIGURE 5.1 – Schème général du processus d'extraction de connaissances

5.2 Phénomènes spatio-temporels considérés

5.2.1 Pollution des rivières

Le réseau hydrographique est un milieu fragile soumis à la présence de nombreuses activités économiques et des usages qui ont modifié, au cours du temps, son intégrité physique et altéré la qualité physico-chimique et biologique de l'eau. Or, les nouvelles réglementations européennes et nationales affichent explicitement la préservation et la restauration des milieux aquatiques et leur environnement. Si des dispositifs de suivi de la qualité de l'eau ont été mis en place depuis plusieurs décennies, il s'agit maintenant de construire des indicateurs permettant de rendre compte de l'influence des usages et des mesures de restauration sur la qualité de l'eau. Pour arriver à la construction de tels outils, il faut prendre en compte différentes types de données : (1) les données hydrologiques, ici, liées à la qualité de l'eau ; (2) les données relatives aux stations de mesure (localisation, réseau d'appartenance, etc.) ; (3) le réseau hydrographique, ses caractéristiques physiques

et les espaces qui lui sont associés : bassin versant, masse d'eau, etc. ; (4) les données relatives aux activités humaines ; et finalement (5) les variables de forçage ou de contexte telles que les données climatiques, ou les données rendant compte de l'homogénéité hydroécologique (comme les hydro-écorégions).

La pollution des rivières, est un phénomène qui est observé en mesurant des indicateurs physico-chimiques et biologiques de la qualité de l'eau. Cette pollution est caractérisée par des évènements qui évoluent au cours du temps et dépendent explicitement de l'emplacement des stations d'échantillonnage situées stratégiquement le long de plusieurs rivières appartenant aux bassins versants. La surveillance de ces stations produit plusieurs données hydrologiques. Les bassins versants sont regroupés dans des entités plus générales associées aux agences de l'eau. Par exemple, l'agence RMC (Rhône-Méditerranée-Corse) gère les bassins versants suivants : le Rhône, la Saône, l'Isère et la Durance. Les données sur lesquelles nous avons travaillé ont été fournies par l'agence de l'eau RMC dans le cadre du projet *Fresqueau*[1]. Ce projet vise à développer de nouvelles techniques d'extraction de connaissances liées à la gestion des données sur la qualité de l'eau. Dans cette thèse, nous nous focalisons sur le bassin versant de la Saône qui est représenté dans la Figure 5.2.

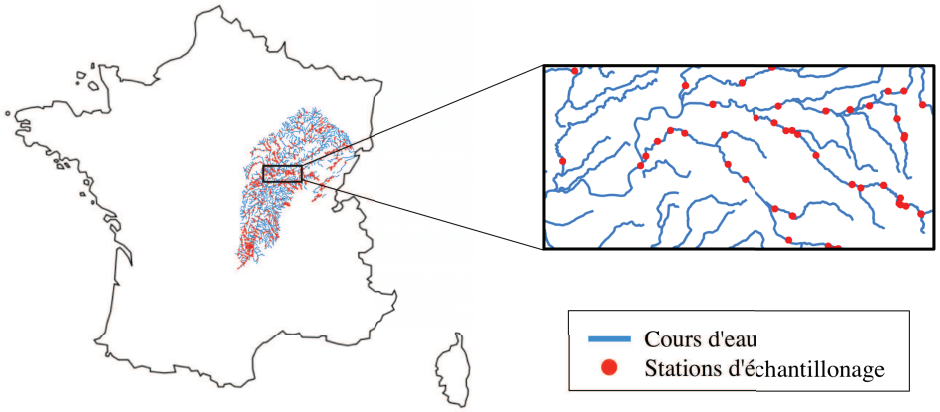


FIGURE 5.2 – Stations d'échantillonnage positionnées le long du bassin versant de la Saône

Les données à notre disposition se présentent sous deux formes : les données statiques qui caractérisent les stations de prélèvement des échantillons et les données dynamiques qui sont issues des prélèvements et permettent, après une analyse, de calculer les indices biologiques. Chaque station d'échantillonnage est identifiée par un code de station *codastace* et est décrite par :

– les *coordonnées Lambert* (x,y) pour identifier la position spatiale de chaque station de prélèvement identifiée par *codstace*. Le système de projection Lambert 93 est ici adopté pour effectuer le géo-référencement ;

1. http://engees-fresqueau.unistra.fr/

- un *point kilométrique* utilisé pour localiser un point le long d'un cours d'eau et qui est calculé en mesurant - en kilomètres - la portion du cours comprise entre le point localisé et un point servant d'origine (la confluence) ;
- une *hydro-écorégion*, qui est une unité spatiale homogène du point de vue de la géologie, du relief et du climat. C'est l'un des principaux critères utilisés dans la typologie et la délimitation de masses d'eau de surface. La France métropolitaine est décomposée en 22 hydroécorégions dont 7 sont présentes dans le bassin versant de la Saône ;
- le *codmasseau*, utilisé pour codifier les masses d'eau correspondant à des eaux superficielles telles qu'une rivière, un canal, ou une partie de rivière, de fleuve ou de canal. Par exemple, pour le bassin versant de la Saône, il existe un total de 572 masses d'eau du type "cours d'eau". En revanche, nous ne traitons pas les masses d'eau tels que les lacs ;
- la *taille de masses d'eau* (très petit, petit, ... , très grand) définie par la position de la masse d'eau dans le réseau hydrographique ;
- un *contexte piscicole*, unité spatiale dans laquelle une population de poissons fonctionne de façon autonome.

Les données dynamiques correspondent aux relevés effectués par les stations d'échantillonnage. La fréquence de ces relevés, ainsi que les informations contrôlées, varient en fonction du temps et des stations. Certaines stations possèdent des relevés récurrents alors que d'autres ne présentent qu'un seul relevé effectué, par exemple, dans le cadre d'une étude ponctuelle.

Les principales informations associées aux relevés sont listées sans le Tableau 5.1.

TABLE 5.1 – Attributs du jeu de données hydrologiques

Attribut	**Description**
idstation	identifiant de la station
rdate	la date de l'échantillon
IBD	Indice Biologique Diatomée. Calcul standardisée du diagnostique de la pollution trophique
ibd2007	la mesure IBD établie avant régulation DCE en France
var_taxo	la variété taxonomique représentant le nombre total de taxons collectés pendant un échantillon, même si lui représente que un seul individu
IBGN	Indice Biologique Global Normalisé. Calcul standardisée basé sur l'identification de macro-invertébrés habitant les rivières
gr_indic	nombre serial du groupe faunistique de l'indicateur IBGN. C'est un indicateur du groupe de taxons appartenant à la même classe de sensibilité à la pollution
index_poiss	indice poisson rivière. Unité spatiale où une population de poisson fonctionne indépendamment

Les mesures *IBGN* et *IBD* ont été traitées sur la base d'une note (e.g. *ibgn_note*) et l'état actuel de la qualité de l'eau (e.g. *ibgn_etat*). En complément, cinq autres variables ont été

incluses dans notre jeu de données : (1) la variété taxonomique représentant le nombre total de taxons collectées au cours d'un prélèvement, même s'ils ne sont représentés que par un seul individu (*var_taxo*) ; (2) le groupe faunistique qui est le plus sensible à la pollution (*gr_indic*) ; (3) l'*indice poison rivière*, qui est une unité spatiale où une population de poisson fonctionne indépendamment ; et (4) la mesure IBD établie avant le règlement DCE en France (*IBD2007*).

Les indicateurs IBGN et IBD sont normalisés en fonction de la masse d'eau étudiée et de l'hydro-écorégion. Trois notes sont alors obtenues et comparables entre les différentes stations : une note pour l'IGBN et une note pour l'IBD et une note correspondant à la fusion normalisée des deux notes précédentes. Au total, 11 attributs appartenant à la dimension d'analyse D_A ont été considérés. Cette dernière information permet d'estimer l'état du cours d'eau au niveau du point de relevé.

Le Tableau 5.2 montre une partie de la base de données associées aux relevés des indicateurs biologiques sur le bassin versant de la Saône.

TABLE 5.2 – Données des relevés biologiques

codstace	**codmasseau**	**x**	**y**	**hydroecor**	**rdate**	**ibgn**	**ibd**	…
6000890	FRDR696	863500	2332140	10	2008-09-23	-100	12	…
6000890	FRDR696	863500	2332140	10	2001-07-10	17	-100	…
6000950	FRDR694	893478	2346387	4	2008-08-28	17	13	…
6000980	FRDR697	866447	2341582	10	2008-08-27	15	12	…
6001250	FRDR691	864725	2323175	10	2003-08-20	-100	12.5	…
6003550	FRDR680	877007	2300933	10	2008-07-31	-100	14	…
6456610	FRDR631	946436	2295348	18	2008-07-19	-100	12.3	…
…	…	…	…	…	…	…	…	…

5.2.2 Suivi épidémiologique de la dengue

La dengue est une infection virale transmise par les moustiques appelés *Aedes aegypti*. Cette infection se produit dans toutes les régions tropicales et subtropicales de la planète. À l'heure actuelle, la dengue est devenue un problème de santé mondiale croissant avec plus des deux cinquièmes de la population mondiale ayant un risque d'infection. La rapide propagation du vecteur de la maladie est attribuée aux changements de démographie, l'urbanisation et l'environnement. La dengue sévère (anciennement connue sous le nom de dengue hémorragique) a été identifiée pour la première fois dans les années cinquante au siècle dernier lors d'une épidémie aux Philippines et en Thaïlande. Aujourd'hui, cette maladie affecte la plupart des pays d'Asie et d'Amérique latine (plus de 100 pays tropicaux et subtropicaux) et est devenue l'une des principales causes d'hospitalisation et de décès chez les enfants de ces régions. Quelques 500 000 personnes infectées pour la dengue sé-

vère nécessitent une hospitalisation chaque année. Une grande proportion d'entre elles sont des enfants et environ $2,5\%$ des personnes touchées par la dengue meurent [2].

Dans le cadre d'une collaboration entre l'Université de la Nouvelle Calédonie, le Ministère des Affaires Sanitaires et Sociales de la Nouvelle Calédonie, l'Institut Pasteur et l'Institut de Recherche pour le Développement, nous avons analysé les données associées à la surveillance épidémiologique de la dengue. Ces données ont été recueillies à Nouméa (Nouvelle Calédonie) sur un territoire divisé en 32 quartiers couvrant $45,7\ \text{km}^2$. Cette division spatiale a été proposée par la Direction des Affaires Sanitaires et Sociales en Nouvelle Calédonie. Cet ensemble de données contient des informations associées à :

- des *données démographiques* : sur la population, liées au recensement et comprennant le code de la zone, le nombre d'habitants, le nombre de maisons, le nombre de ménages, etc. Ces données ont été recueillies au cours de deux années (1996 et 2004) pour chaque quartier de Nouméa ;
- *des données entomologiques* : associées à des caractéristiques de la transmission du vecteur *Aedes aegypti*. Ces données comprennent l'Indice Breteau (IB), l'indice entomologique de haut risque (IHRE), des données relatives au sérotype, le taux de positivité et d'autres informations utilisées par les épidémiologistes pour analyser la maladie. Dans cette catégorie, les données ont été récupérées par zone et la granularité temporelle est mensuelle ;
- des *données météorologiques* : associées à des informations météorologiques telles que les précipitations (mm), la force du vent (m/s), la température (°C), l'humidité (%), entre autres. Ces données ont été enregistrées quotidiennement par deux capteurs installés à Nouméa ;
- des *données de planification urbaine* : les moustiques *Aedes aegypti* vivent et se reproduisent dans les zones urbaines à proximité de l'homme. Les moustiques se développent dans des récipients artificiels qui servent de dépôts d'eau (par exemple, de vieux pneus, des plateaux de plantes en pot, des urnes ou des conteneurs de l'eau de pluie) et se nourrissent généralement de sang humain. Tenant compte de ces deux aspects, nous intégrons : (1) d'un côté, les données associées aux endroits essentiels pour le développement des moustiques comme le nombre de piscines, le nombre de serres, le nombre de bassins, des fontaines ; et (2) les données relatives aux endroits où les gens se réunissent pour des activités sociales, par exemple, les écoles, les églises, les crèches, etc.
- des *données médicales* : ces données ne comprennent que le nombre de cas de dengue signalés à Nouméa par jour. Comme on peut l'observer sur la Figure 5.3, les années 1996 et 2003 sont les plus importantes en termes de nombre de cas de dengue. Nous avons utilisé uniquement les données de 2003.

Les attributs du jeu de données de la dengue pour l'année 2003 sont décrits dans le Tableau 5.3.

2. Site official de l'Organisation Mondiale de la Santé.

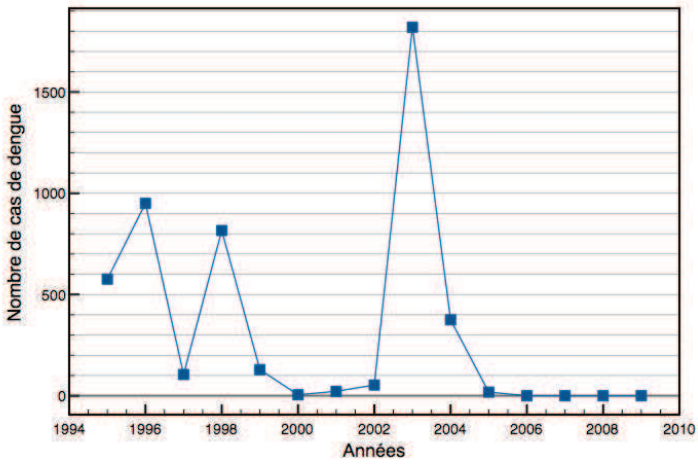


FIGURE 5.3 – Nombre de cas de dengue par année à Nouméa

TABLE 5.3 – Attributs du jeu de données de la dengue

Attribut	**Description**
id_quartier	ID de la zone (quartier)
date	date de l'échantillon
precip	précipitation en mm par quartier
mean_wind	force du vent moyenne en m/s par quartier
mean_temper	température moyenne en °C par quartier
mean_humid	humidité moyenne en % par quartier
outdoor_deposit	nombre de points d'eau dans des zones publiques (étangs, rivières, fontaines, etc.)
graveyard	nombre de cimetières par quartier
waste_container	nombre de poubelles par quartier
indoor_deposit	nombre de points d'eau d'ordre privatif (drainages, conduits, puits, etc.)
community_gather	nombre de centres communales (écoles, églises, universités, etc.)
ihre_index	indice d'haut risque entomologique
nb_cas_dengue	nombre de cas de dengue par quartier

Le Tableau 5.4 montre une partie de la base de données associées aux données de suivi épidémiologique de la dengue en Nouvelle Calédonie.

5.2.3 Discrétisation des jeux de données réelles

L'étude de la répartition des valeurs pour l'ensemble des attributs, nous a permis de définir une méthode de discrétisation. Pour les données mises à notre disposition (données

TABLE 5.4 – Données associées à la dengue

id quartier	population	the_geom	date	température	humidité	nombre de cas de dengue	...
Q12	419	010600001...	2003-09-23	22	63	2	...
Q12	419	010600001...	2003-09-26	21	63	1	...
Q2	3343	010600001...	2006-01-03	23	78	18	...
Q18	2409	010600001...	2008-12-11	22	65	12	...
Q9	3487	010600001...	2009-06-23	20	60	15	...
Q11	9775	010600001...	2006-04-05	22	65	9	...
...	...	...	...	...	...	...	...

hydrographiques et données de la dengue), les composantes sont bien séparées, le nombre d'observations suffisant et la seule considération de l'histogramme de fréquences fourni une estimation correcte du nombre de constituants et de leurs valeurs. Afin d'obtenir des données catégorielles [3], une discrétisation a été faite pour transformer des données continues en données nominales en utilisant la technique des *fréquences égales* (*equi-largeur binning* en anglais). Pour cela, nous avons construit une courbe de fréquences accumulées pour déterminer les limites de chaque classe en fonction des seuils observés sur la courbe. Les plages de valeurs sont choisies pour égaliser la distribution des valeurs de chaque variable. Avec ce type de discrétisation, nous obtenons des plages de valeurs équilibrées selon trois classes pour les données de la dengue et selon quatre classes pour le jeu de données hydrographiques.

5.2.4 Générateur de données spatio-temporelles synthétiques

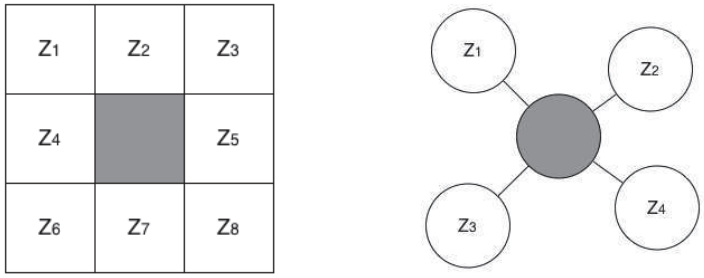
Les données synthétiques sont définies comme des *données applicables à une situation quelconque et qui ne sont pas obtenues par des mesures directes* [4]. Pour des raisons évidentes, il est souhaitable que les données générées synthétiquement soient plus ou moins réalistes et reflètent les propriétés essentielles de l'ensemble des données sur lesquelles on souhaite travailler. Ces jeux synthétiques doivent également permettre de tester des cas limites [Eno et Thompson, 2008].

Pour obtenir des données spatio-temporelles générées aléatoirement ainsi qu'une relation de voisinage associée à ces données, nous avons développé un générateur de données synthétiques, adaptable selon les paramètres suivants : le nombre de zones, le type de relation de voisinage, le nombre de dates par zone et le nombre moyen d'items par zone. Ce générateur permet de construire deux types de relations de voisinage (voir Figure 5.4).

3. Données séparées dans des catégories qui s'excluent mutuellement.

4. McGraw-Hill Science & Technology Dictionary McGraw-Hill Dictionary of Scientific and Technical Terms. Copyright 2003.

La première est utilisée pour représenter l'espace comme une grille dans laquelle chaque carré représente une région, une zone sera entourée par huit voisins. La seconde est utilisée pour représenter l'espace comme un graphe dont les sommets et les arêtes représentent les zones et leur relation de voisinage respectivement. Dans ce dernier cas, le générateur prend comme paramètres d'entrée le nombre d'arêtes et le nombre de sommets. Différents jeux de données synthétiques ont été générés pour évaluer la performance de nos algorithmes.



(a) (b)

FIGURE 5.4 – Types de relation de voisinage (a) grille (b) graphe

5.3 Extraction de motifs spatialement fréquents

Dans cette section, nous allons présenter l'application de la première approche d'extraction de motifs spatialement fréquents (cf., Section 3.2) sur le jeu de données associées à la pollution des rivières comportant le bassin versant de la Saône et sur le jeu de données associées au suivi épidémiologique de la dengue en Nouvelle Calédonie. Cette deuxième application sera moins détaillée mais nous permettra de démontrer la généralité de notre approche.

5.3.1 Extraction de motifs spatialement fréquents sur les données associées à la pollution des rivières

Les données dont nous disposons sont géo-référencées et variables dans le temps, ce qui les rend difficile à explorer conjointement. D'ailleurs, la relation spatiale entre les entités étudiées (stations de surveillance) est implicite (e.g. une station est située en amont d'une autre ou proche en distance mais sur des cours d'eau différents). Il est donc nécessaire d'effectuer des pré-traitements permettant de prendre en compte différentes proximités spatiales (par le regroupement des stations en fonction de leur distance, ou de leur appartenance à un cours d'eau, etc.) afin de construire des zones homogènes.

De façon à prendre en compte les caractéristiques spatiales des données, nous proposons trois approches de spatialisation détaillées dans la suite de cette section.

Division de l'espace

Nous souhaitons explorer les données de qualité de l'eau (ici des indicateurs biologiques) en prenant en compte les informations spatiales (e.g. localisation des stations le long des rivières, proximité, etc.) tout en nous préoccupant de l'évolution temporelle des données considérées. L'approche que nous adoptons est composée de deux phases : (1) la préparation des données et la division de l'espace ; et (2) l'extraction de motifs prenant en compte l'aspect temporel des données. Ce processus général est illustré sur la Figure 3.4. De façon plus détaillée, une série de pré-traitements traduit la prise en compte de différentes proximités spatiales (e.g. rapprochement des stations selon leur distance, selon leur appartenance à une zone, etc.). Pour retracer la prise en compte de l'historique des données, nous appliquons une méthode de recherche de motifs séquentiels définis par Agrawal et Srikant [1995]. Ces motifs correspondent à des séquences fréquentes d'états mesurés par les stations.

Les données spatiales nous permettent de déterminer des zones géographiques pertinentes afin de gérer : (1) la proximité à partir de la localisation des stations exprimée par leurs coordonnées Lambert (système de coordonnées géo-référencées) ; et (2) la proximité liée au cours d'eau à partir du sens de circulation des écoulements et des connexions entre cours d'eau.

Nous explorons ainsi les données de deux manières différentes dans le but de regrouper les stations de surveillance en zones homogènes. Cela permet d'étudier comment les évènements apparues dans des zones voisines peuvent avoir des répercussions dans les zones étudiées.

Dans cette thèse, en plus de l'approche naïve (sans découpage en zones), deux découpages de l'espace ont été proposés en suivant les hypothèses définies dans la Section 3.2.3 :

1. Un voisinage restreint au *cours d'eau* : pour un cours d'eau quelconque, deux stations X et Y positionnées dans ce cours d'eau sont considérées comme voisines. Par exemple, dans la Figure 5.5, les stations de W, X, Y et Z appartiennent au même cours d'eau, donc, ces stations sont considérées comme une seule zone et leurs données sont combinées. Un exemple d'incident que nous pouvons étudier grâce à cette approche est : l'écoulement de carburant d'un bateau autour de la station X aura un impact sur les mesures, d'abord sur la station X et plus tard sur les mesures des stations Y et Z situées en aval de la station X.
2. Le ϵ-*voisinage* : l'espace est divisé en zones regroupant des stations en exploitant les coordonnées Lambert. Dans chacune de ces zones, les stations situées dans une superficie couvrant ϵ km^2 sont regroupées, même si ces stations appartiennent à différents cours d'eau. Par exemple, dans la Figure 5.6, les stations de W, X et Y sont

FIGURE 5.5 – Division de l'espace en utilisant l'approche *cours d'eau*

considérées comme une seule zone, même si elles ne sont pas sur le même cours d'eau. Un exemple de phénomène que nous pouvons étudier grâce à cette approche est : l'utilisation de pesticides dans un champ de culture situé entre les stations X et Y peut avoir un impact sur les mesures des stations situées sur les rivières environnantes à cette zone de culture, même si ces stations ne se trouvent pas dans la même rivière.

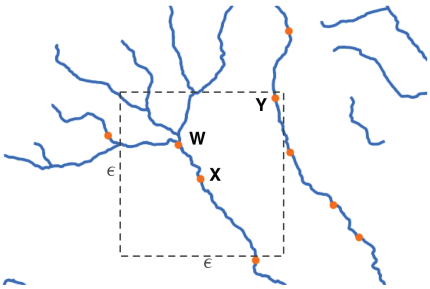


FIGURE 5.6 – Division de l'espace en utilisant l'approche ϵ*-voisinage*

Ces deux découpages vont mettre en évidence deux hypothèses différentes pour l'influence de la pollution : (1) la première hypothèse est que la pollution mesurée dans un cours d'eau donné à une station particulière X, va d'une part avoir potentiellement un impact sur les stations situées en aval de X et d'autre part, que l'origine de la pollution est liée à un phénomène situé en amont de X. Le découpage du type *cours d'eau* nous permet donc d'évaluer cette hypothèse en "moyennant" les indicateurs de pollution à l'intégralité d'un cours d'eau ; et (2) la seconde hypothèse est que la pollution mesurée dans une station X est le résultat d'une pollution dont l'origine peut se situer sur le même cours d'eau, dans des nappes souterraines, dans des zones agricoles voisines, etc. Le découpage appelé ϵ*-voisinage* nous permet donc de "moyenner" les indicateurs de pollution dans des zones suffisamment grandes pour observer des effets potentiellement non limités au cours d'eau associé à la station X.

Fouille de motifs spatialement fréquents

Les expérimentations ont été réalisées à partir de la base de données d'indicateurs biologiques relevés dans les rivières de la Saône (cf., Tableau 5.2).

Des problèmes récurrents dans l'extraction de motifs fréquents sont l'hétérogénéité et l'absence de certaines valeurs dans les données. En effet, le jeu de données du basin versant de la Saône contient un nombre important de valeurs manquantes non considérées dans nos tests, pour éviter d'extraire des motifs du type ⟨(ibgn_note_2, NULL, NULL, NULL, NULL)⟩.

Nous avons extrait des séquences de motifs selon les trois approches de spatialisation.

1. En considérant l'espace comme une seule unité, c'est-à-dire, sans le découper en zones (*NZ*). Nous avons donc considéré 711 stations de prélèvement identifiées par le code de la station.
2. En utilisant le voisinage du type *cours d'eau*, pour diviser l'espace en zones plus ou mois hétérogènes : le découpage de l'espace en utilisant ce type de voisinage nous a fourni un total de 233 zones.
3. En considérant le ϵ-*voisinage*, où ϵ a été fixé dans un premier temps à 10, définissant ainsi des zones de 100 km^2. Nous avons ainsi obtenu 223 zones.

Pour extraire des motifs spatialement fréquents, nous avons utilisé l'algorithme *PrefixSpan* [Mortazavi-Asl *et al.*, 2000] car il a été démontré, dans de nombreux articles, que cet algorithme est très efficace pour fouiller de grands volumes de données. Le principe de cette approche est d'extraire des motifs fréquents sans la génération de motifs candidats, comme dans les approches dites *Apriori*. Pour nos expérimentations, nous avons utilisé le prototype appelé *Sequential Pattern Mining Frameworf (SPMF)*[5] sur les trois jeu de données cités.

Les premiers résultats des expérimentations appliquées aux données avec les trois approches de spatialisation proposées sont :

1. Pour l'approche *NZ*, pour un support minimum de 0,3, nous avons obtenu 22 motifs fréquents, tous de taille 1. Le Tableau 5.5 montre quelques motifs extraits ;

TABLE 5.5 – Exemple de motifs obtenus pour l'approche de spatialisation *NZ*

Motifs	**Support**
⟨(ibgn_etat_TBE)⟩	0,32
⟨(ibgn_etat_TBE, ibgn_note_4)⟩	0,32
⟨(ibgn_0-10 gr_indic_0-4)⟩	0,32
⟨(ibgn_etat_BE ibgn_note_3)⟩	0,31
...	...

5. Disponible sur http://www.philippe-fournier-viger.com/spmf/

2. Pour la deuxième approche *cours d'eau*, avec un support minimum de 0,3, nous avons obtenu 564 motifs fréquents, parmi lesquels 110 sont de taille 1, 361 sont de taille 2, 90 de taille 3 et 3 de taille 4. Une partie des résultats trouvés est présentée dans le Tableau 5.6.

TABLE 5.6 – Exemple de motifs obtenus pour l'approche de spatialisation *cours d'eaux*

Motifs	**Support**
⟨(ibgn_11-15)(ibgn_11-15 var_taxo_21-30)⟩	0,41
⟨(var_taxo_21-30)(var_taxo_21-30)(ibgn_11-15)⟩	0,36
⟨(ibgn_11-15 ibgn_etat_Emoy ibgn_note_2)(ibgn_11-15)⟩	0,35
⟨(ibgn_note_2)(ibgn_etat_Emoy ibgn_note_2)⟩	0,31
⟨(gr_indic_5-6 var_taxo_21-30 ibgn_etat_Emoy ibgn_note_2)⟩	0,30
⟨(var_taxo_21-30)(ibgn_11-15 var_taxo_21-30)(var_taxo_21-30)⟩	0,33
⟨(var_taxo_21-30)(ibgn_16-20 var_taxo_31-40)⟩	0,30
…	…

3. Pour la troisième approche ϵ*-voisinage*, avec un support minimum de 0,3, nous avons obtenu 138 motifs fréquents de taille 1, 1 174 motifs fréquents de taille 2, 658 de taille 3, 104 de taille 4 et 8 motifs de taille 5. Au total, 2 082 motifs fréquents sont extraits. Une partie des résultats trouvés est présentée dans le Tableau 5.7.

TABLE 5.7 – Exemple de motifs obtenus pour l'approche de spatialisation ϵ*-voisinage*

Motifs	**Support**
⟨(ibgn_etat_Emoy ibgn_note_2)(var_taxo_21-30)⟩	0,42
⟨(var_taxo_21-30)(var_taxo_21-30)(ibgn_11-15)⟩	0,42
⟨(ibgn_note_2)(ibgn_etat_Emoy ibgn_note_2)⟩	0,36
⟨(gr_indic_7-9)(ibgn_16-20 gr_indic_7-9 ibgn_etat_TBE, ibgn_note_4)⟩	0,39
⟨(var_taxo_21-30 ibgn_note_2)(ibgn_11-15 var_taxo_21-30)⟩	0,35
⟨(var_taxo_21-30 ibgn_etat_Emoy)(var_taxo_21-30)(ibgn_11-15)⟩	0,31
⟨(var_taxo_21-30 ibgn_etat_Emoy ibgn_note_2)(ibgn_11-15)(ibgn_11-15 var_taxo_21-30)⟩	0,31
…	…

Dans les Tableaux 5.5, 5.6 et 5.7, chaque séquence est composée d'une liste d'évènements associés à une estampille temporelle (représentée par des parenthèses). Par exemple, dans le Tableau 5.7, l'itemset (var_taxo_21-30 ibgn_etat_Emoy) représente la valeur entre 21 et 30 de l'attribut *variable taxonomique* et une valeur moyenne de l'attribut *état de l'ibgn* apparaissant conjointement dans la même estampille temporelle. L'écart temporel entre deux itemsets est représenté à titre indicatif et seulement l'ordre importe.

Le dernier motif du Tableau 5.6 peut être interprété comme : dans 30% des cours d'eau, la *variété taxonomique* augmente au cours du temps. Nous pouvons attribuer cette augmentation à la présence de l'évènement *ibgn_16-20* apparu dans la deuxième estampille temporelle (voir Figure 5.7).

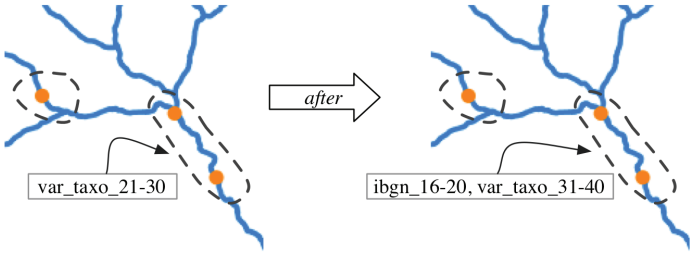


FIGURE 5.7 – Interprétation graphique de la séquence ⟨(var_taxo_21-30)(ibgn_16-20 var_taxo_31-40)⟩

Le nombre de motifs obtenus lors de l'exécution de l'algorithme PrefixSpan sur le jeu de données avec les trois approches de spatialisation et un support minimum de 0,3 est respectivement 22 pour l'approche *NZ*, 564 pour *cours d'eau* et 2 082 pour ϵ*-voisinage*. Il est intéressant de constater que nous avons extrait peu de motifs fréquents en utilisant l'approche de spatialisation naïve par rapport aux deux autres approches. De plus, les motifs extraits pour cette première approche sont de taille 1. Ces motifs ne capturent pas l'évolution temporelle des données, à la différence des motifs extraits en utilisant les données soumises au pré-traitement de spatialisation.

Extraction de motifs spatialement fréquents peu contredits

Nous avons appliqué une mesure objective [6] de validation sur les motifs extraits afin de filtrer les plus pertinents pour les experts. En effet, même si en terme de volume une validation exhaustive est envisageable, cela ne le sera plus dès lorsque le jeu de données sera étendu au niveau national. La mesure de qualité utilisée dans cette partie expérimentale est la *moindre contradiction temporelle (MCT)*, définie dans la Section 4.3.

La MTC a été calculée de la façon suivante : Soit $\mathfrak{mBD}$ la base de motifs spatialement fréquents obtenus après l'exécution de l'algorithme PrefixSpan sur le jeu de données du bassin versant de la Saône en considérant, par exemple, la spatialisation du ϵ*-voisinage*. Soit un motif $s \in \mathfrak{mBD}$ et son support $supp(s)$ décrits dans le Tableau 5.8.

TABLE 5.8 – Séquence s

Motif	**Support**
⟨(ibgn_16-20 ibgn_etat_TBE)(var_taxo_31-40)⟩	0,34

Pour le calcul de S_{contr}, nous cherchons les itemsets (ibgn_1620 ibgn_etat_TBE) et (var_taxo_31-40) dans toutes les séquences s_i de la base de données sans considérer la

6. Non dépendante d'un point de vue mais associée à un calcul.

position d'apparition des itemsets étudiés dans s_i. Nous les trouvons deux fois (voir Tableau 5.9).

TABLE 5.9 – Séquences appartenant à l'ensemble S_{contr}

Motifs	Support
⟨**(ibgn_16-20 ibgn_etat_TBE)**(ibgn_11-15)**(var_taxo_31-40)**⟩	0,34
⟨**(var_taxo_31-40)(ibgn_16-20 ibgn_etat_TBE)**⟩	0,32

Finalement, la valeur S_{contr} pour la séquence ⟨(ibgn_16-20 ibgn_etat_TBE) (var_taxo_31-40)⟩ est 0,66.

Le calcul de S_{all} se fait de façon analogue au calcul de S_{contr}. Nous cherchons les évènements (items) appartenant à la séquence ⟨(ibgn_16-20 ibgn_etat_TBE)(var_taxo_31-40)⟩ dans toutes les séquences s_i de la base de motifs mBD. Nous les trouvons dans les motifs montrés dans le Tableau 5.10. La somme des supports des séquences $s_a \in S_{all}$ est égale à 3,34.

TABLE 5.10 – Séquences appartenant à l'ensemble S_{all}

Motifs	Support
⟨(**ibgn_16-20 var_taxo_31-40 ibgn_etat_TBE**)⟩	0,36
⟨(**ibgn_16-20 var_taxo_31-40 ibgn_etat_TBE** ibgn_note_4)⟩	0,36
⟨(**ibgn_16-20** gr_indic_7-9 **var_taxo_31-40 ibgn_etat_TBE**)⟩	0,34
⟨(**ibgn_16-20** gr_indic_7-9 **var_taxo_31-40 ibgn_etat_TBE** ibgn_note_4)⟩	0,34
…	…

Finalement, la moindre contradiction temporelle (MCT) pour la séquence s =⟨(ibgn_16-20 ibgn_etat_TBE)(var_taxo_31-40)⟩ est :

$$\begin{aligned} MCT(s) &= \frac{0,34-0,66}{3,34} \\ &= -0,0958 \end{aligned}$$

La mesure objective de validation MCT a été appliquée aux motifs obtenus lors de l'exécution de l'algorithme choisi sur le jeu de données des bassins versants de la Saône pour les trois approches de spatialisation proposées.

Les Tableaux 5.11, 5.12 et 5.13 montrent quelques motifs, leur support et la valeur de la moindre contradiction temporelle MCT pour les différentes approches de spatialisation.

Le Tableau 5.13 montre des motifs, leur support ainsi que la moindre contradiction temporelle associée aux motifs obtenus en utilisant l'approche de spatialisation ϵ-*voisins*. Nous pouvons noter, par exemple, que le motif ⟨(ibgn_11-15)(ibgn_16-20, gr_indic_7-9)⟩ apparaît dans 33% des zones et n'est pas contredit par les données au cours du temps.

TABLE 5.11 – MCT pour les motifs extraits sur les données sans spatialisation *NZ*

Motifs	**Support**	**MCT**
⟨(ibgn_etat_TBE, ibgn_note_4)⟩	0,32	1,0
⟨(ibgn_11-15 var_taxo_21-30)⟩	0,39	1,0
⟨(var_taxo_21-30)⟩	0,5	0,1236
⟨(ibgn_0-10)⟩	0,36	0,05882
…	…	…

TABLE 5.12 – MCT pour les motifs extraits sur les données en utilisant l'approche *cours d'eau*

Motifs	**Support**	**MCT**
⟨(var_taxo_21-30)(ibgn_16-20 var_taxo_31-40)⟩	0,3	1,0
⟨(ibgn_0-10 gr_indic_0-4, ibgn_etat_Emedio ibgn_note_1)⟩	0,32	1,0
⟨(ibgn_0-10 ibgn_etat_Emedio ibgn_note_1)⟩	0,34	0,0303
⟨(ibgn_note_1)⟩	0,35	−0,738806
⟨(ibgn_note_4) (ibgn_etat_TBE)⟩	0,34	−0,963176
…	…	…

TABLE 5.13 – MCT pour les motifs extraits sur les données en utilisant l'approche ϵ-*voisinage*

Motifs	**Support**	**MCT**
⟨(ibgn_11-15)(ibgn_16-20 gr_indic_7-9)⟩	0,33	1,0
⟨(var_taxo_21-30)(ibgn_etat_TBE ibgn_note_4)⟩	0,33	0,03125
⟨(gr_indic_7-9)(ibgn_11-15 var_taxo_21-30)⟩	0,36	0,01887
⟨(gr_indic_7-9)(ibgn_etat_BE ibgn_note_3)⟩	0,31	−0,030928
⟨(var_taxo_21-30)(var_taxo_31-40)⟩	0,42	−0,215329
⟨(ibgn_etat_TBE ibgn_note_4)(var_taxo_31-40 ibgn_note_4)⟩	0,33	−0,905918
…	…	…

Contrairement, le motif ⟨(var_taxo_21-30)(ibgn_etat_TBE, ibgn_note_4)⟩ apparaît aussi dans 33% des zones mais est très contredit. Nous pouvons conclure que le premier motif est plus intéressant pour les experts.

Pour comparer la similarité entre les motifs obtenus, nous avons utilisé la mesure S^2MP proposée par Saneifar *et al.* [2008] pour identifier les irrégularités et construire des classes d'objets homogènes. Nous avons appliqué l'approche avant citée aux motifs obtenus pour les deux spatialisations : *cours d'eau* et ϵ-*voisinage*. Nous n'avons pas appliqué la méthode aux motifs obtenus en utilisant l'approche de spatialisation naïve *NZ* car le nombre de motifs est très réduit. Saneifar *et al.* [2008] ont proposé une mesure permettant de comparer deux séquences au niveau des itemsets et de leur position dans la séquence ainsi qu'au

niveau de la ressemblance des items dans les itemsets. Les auteurs ont aussi étendus la technique de clustering k-*means* pour les séquences d'itemsets.

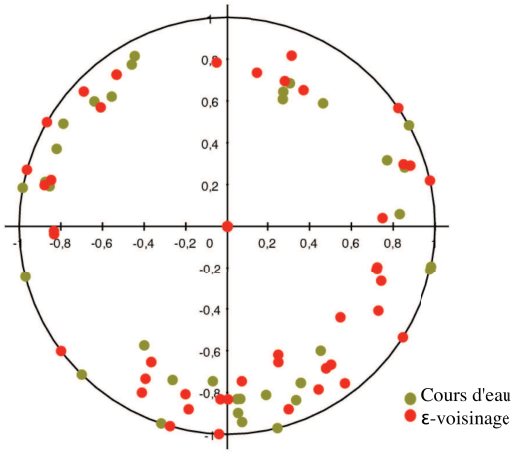


FIGURE 5.8 – Déploiement des motifs séquentiels regroupés par la distance autour de *centroids* pour les deux approches de spatialisation

Pour identifier quelle approche de spatialisation est la plus pertinente pour le jeu de données du bassin versant de la Saône, nous avons regroupé les clusters obtenus pour les deux jeux de motifs associés aux deux approches de spatialisation. La Figure 5.8 montre justement les clusters dans un seul plan autour d'un *centroide*. Dans cette figure, nous pouvons noter que les motifs obtenus en utilisant l'approche de spatialisation *cours d'eau* sont légèrement plus proches du centre du plan que les points représentant les motifs extraits avec ϵ-*voisinage*. Cette différence peut être confirmée grâce à la fonction objective appelée *sum of square errors* (SSE). La valeur du SEE, pour un cluster donné, est calculée de la façon suivante : pour chaque instance du cluster, sommer les carrés de la différence entre chaque valeur de l'attribut de chaque objet et la valeur du centroide du cluster auquel il appartient. Ces valeurs sont additionnées pour chaque instance du cluster et pour tous les clusters. La valeur du SSE, qui représente la cohésion des clusters, est égale à 22 4136 en utilisant l'approche de spatialisation *cours d'eau*, qui est inférieure à la valeur du SSE obtenue pour les clusters utilisant l'approche ϵ-*voisinage* (autour de 25 3068). La cohésion entre les instances en utilisant l'approche *cours d'eau* étant plus forte que l'autre approche, elle est par conséquence plus intéressante pour les experts.

5.3.2 Extraction de motifs spatialement fréquents sur les données associées au suivi épidémiologique de la dengue

Le dengue est une maladie transmise pour le vecteur appelé *Aedes aegypti*, qui a un mode de vie qui le rend particulièrement proche de l'homme. Diamond [1998] affirme que

les épidémies sont en relation directe avec des populations nombreuses et denses et souvent vivant en état de précarité. Dans ce contexte, le *modus vivendi* des humains peut avoir un impact sur le nombre de cas de dengue enregistrés dans une région ou un quartier. Les lieux associés au groupements de gens (e.g. écoles, zones urbaines, etc.) sont plus sensibles à une possible épidémie que les zones d'aménagements associées aux espaces verts non utilisées par les humains.

Par ailleurs, les risques environnementaux associés aux activités humaines, tels que les loisirs, l'industrie, etc. sur le comportement des vecteurs de transmission sont importants. Par exemple, l'utilisation de produits chimiques dans des zones industrielles et/ou commerciales peut avoir un impact considérable dans la prolifération (ou le contraire) des moustiques, vecteurs de la dengue.

Division de l'espace

Les facteurs décrits précédemment, doivent être pris en compte au moment de l'extraction des connaissances sur des données associées au suivi épidémiologique de la dengue. L'idée ici est d'étudier l'impact des régions géographiques associées aux activités humaines sur la propagation de la maladie.

Pour cela, nous proposons deux approches de spatialisation (pré-traitement) dans l'intention d'étudier l'évolution des évènement décrivant une zone au cours du temps en prenant en compte deux granularités spatiales : (1) au niveau du *quartier* i.e., chaque quartier de la ville de Nouméa (entité spatiale) sera étudié indépendamment ; et (2) au niveau des *zones d'aménagements* spécifiques, i.e., nous allons regrouper les quartiers par zone représentant diverses activités humaines comme les loisirs, l'industrie, etc.

Concernant la première approche de spatialisation, le découpage en quartiers de la ville de Nouméa a été proposé par la mairie de Nouméa. Il sera utilisé au cours de nos expérimentations (voir Annexe A). Par ailleurs, les données associées au suivi épidémiologique de la dengue dont nous disposons ont une granularité spatiale au niveau du quartier. Nouméa est divisé en 32 quartiers dont 12 seront étudiés dans cette section car ils comportent au moins un cas de dengue par mois pour l'année 2003 (l'année dans laquelle il y a eu une forte épidémie).

Par rapport à la deuxième approche de spatialisation, elle est plus complexe à mettre en place car l'agglomération de Nouméa pose de multiples problèmes en termes d'organisation et de gestion de l'espace et l'on constate un fort déséquilibre humain et économique. Le découpage urbain en zones homogènes conditionne le cadre résidentiel et l'accès aux biens et services auxquels les habitants font appel. Dans ce processus, souvent appelé *ségrégation*, nous procédons à un découpage du territoire en secteurs géographiques homogènes et équilibrés en prenant en compte des caractéristiques spécifiques d'une entité spatiale appartenant à une zone (e.g. une vallée) ou associé à une activité (e.g. l'industrie).

À cet égard, nous allons regrouper des quartiers constituant des secteurs. Chaque secteur représente une activité : zones urbaines, zones industrielles, espaces verts aménagés

et non aménagés, etc. Cette répartition des espaces a été proposée par *Yves Mermoud*[7]. Dans cette thèse, cinq zones seront prise en compte : (1) tissu urbain dense ; (2) habitat individuel dominant ; (3) zone touristique et de loisirs ; (4) zone industrielle ; et (5) zone portuaire et militaire.

Les figures présentées dans l'Annexe A décrivent cette décomposition en zones homogènes.

Fouille de motifs spatialement fréquents

Les expérimentations ont été réalisées à partir d'une base de données associées au suivi épidémiologique de la dengue décrit dans la Section 5.2.2. Ce jeu de données est constitué de 11 attributs d'analyse. Chaque registre de la base de données est associé aussi à un identifiant du quartier et à la date d'enregistrement.

Nous avons pris en compte une granularité temporelle fixée à une semaine car le vecteur de transmission de la dengue ne vit pas plus de 15 jours. Il est donc important de considérer cette contrainte temporelle lors de nos expérimentations.

Nous avons extrait des motifs séquentiels selon deux approches de spatialisation décrites dans la Section 5.3.2) :

1. Une approche *naïve* : nous ne regroupons pas les entités spatiales, c'est-à-dire, chaque quartier est une entité spatiale. Nous avons au total 12 entités.
2. En utilisant l'approche de spatialisation en *zones d'aménagements*, pour diviser l'espace en zones plus ou mois hétérogènes. Le groupement des quartiers en utilisant cette approche de spatialisation nous a fourni un total de 5 zones.

Le Tableau 5.14 présente un résumé des deux caractéristiques associées aux deux jeux de données construits selon les deux approches de spatialisation proposées précédemment. Le nombre de dates représente le nombre moyen de semaines existant dans les mois appartenant à la période épidémique (très fréquemment de Janvier jusqu'à Juin). Le nombre d'items représente le nombre d'attributs de la dimension d'analyse D_A qui pour toutes les estampilles temporelles est fixé à 11.

TABLE 5.14 – Caractéristiques des jeux de données associées à la dengue

Approche de spatialisation	Nombre de zones	Nombre de dates par zone (moyenne)	Nombre d'items par date	Granularité temporelle	Granularité spatiale
Par quartier	12	24	11	hebdomadaire	quartiers
Par zones d'aménagements	5	24	11	hebdomadaire	zone d'aménagements

Suivant le même protocole d'expérimentation proposé dans la section précédente, nous avons utilisé l'implémentation de l'algorithme *PrefixSpan* sur les jeux de données construits en utilisant les deux approches de spatialisation.

7. Enseignant au lycée Lapérouse, Nouvelle Calédonie http://www.ac-noumea.nc/laperouse/

Les premiers résultats des expérimentations sont :

1. Pour l'approche *naïve*, pour un support minimum de 0,5, nous avons obtenu 797 motifs fréquents. Le Tableau 5.15 montre quelques motifs extraits ;

TABLE 5.15 – Exemple de motifs obtenus pour l'approche de spatialisation *naïve*

Motifs	**Support**
⟨(nb_cas_dengue :(4.00 ;6.00])(mean_humid :(76.85 ;83.75] positif_aedes :<=1.00)⟩	0,67
⟨(mean_temper :<=23.45)(mean_humid :<=76.85 nb_cas_dengue :(4.00 ;6.00])⟩	0,50
⟨(community_gather :<=17.50 ihre_index :<=1.00 nb_cas_dengue :(4.00 ;6.00])⟩	0,50
…	…

2. Pour la deuxième approche *zones d'aménagements*, avec un support minimum de 0,7, nous avons obtenu 1 037 motifs fréquents. Une partie des résultats trouvés est présentée dans le Tableau 5.16.

TABLE 5.16 – Exemple de motifs obtenus pour l'approche de spatialisation *zones d'aménagements*

Motifs	**Support**
⟨(mean_wind :<=3.29)(mean_temper :>23.32 mean_humid :>83.10) (mean_temper :<=23.32)(nb_cas_dengue :<=4.00)⟩	1,00
⟨(mean_temper :>23.32 nb_cas_dengue :(4.00 ;5.00]) (mean_wind :<=3.29 nb_cas_dengue :<=4.00)⟩	1,00
⟨(mean_temper :>23.32 ihre_index :<=1.00) (mean_humid :<=77.15)(nb_cas_dengue :<=4.00)⟩	0,80
…	…

Dans ces expérimentations, nous pouvons constater que le nombre de séquences obtenues lors de l'exécution de l'algorithme *PrefixSpan* sur le jeu de données avec les deux approches de spatialisation et un support minimum de 0,7 est respectivement 10 pour l'approche *naïve* et 1 037 pour l'approche *zones d'aménagements*. Cet effet a été constaté aussi dans les expérimentations précédentes (sur le jeu de données associées à la pollution des rivières).

À titre d'exemple, le motif ⟨(comunity_gather :<=17.50 ihre_index :<=1.00 nb_cas_dengue :(4.00 ;6.00])⟩ du Tableau 5.15 indique que dans la moitié des quartiers, la présence de dengue est associée aux évènements *community_gather* et *ihre_index*. Ce dernier comporte, entre autres, la présence de gites larvaires du vecteur de transmission de la maladie.

Le deuxième exemple du Tableau 5.16 peut être interprété comme : dans toutes les zones d'aménagements étudiées, il y a une diminution du nombre de cas de dengue. Cette

diminution peut être mise en relation avec l'évènement *mean_wind :<=3.29* apparu à la deuxième estampille temporelle.

Nous pouvons constater aussi, que les exemples montrés dans le Tableau 5.16, les évènements *graveyard*, *community_gather* associés aux rassemblements de personnes n'apparaissent pas dans les résultats. En effet, avec un support proche de 1, seule les évènements associés aux phénomènes météorologiques sont extraits. Par exemple, dans les zones d'aménagement associées aux espaces verts, les évènements tels que *community_gather* apparaissent rarement.

Extraction de motifs spatialement fréquents peu contredites

Nous avons appliqué la *moindre contradiction temporelle* sur les motifs extraits dans l'étape précédente afin de filtrer les plus pertinents pour les experts. Les Tableaux 5.17 et 5.18 montrent les motifs extraits, leur support ainsi que la valeur associée à la *moindre contradiction temporelle*.

TABLE 5.17 – MCT pour les données sans zonage *naïve*

Motifs	**Support**	**MCT**
⟨(mean_temper :<=23.45 outdoor_deposit :<=122.50 waste_container :<=29.50)⟩	0,70	0,982
⟨(graveyard :<=0.50 indoor_deposit :<=2398.00 nb_cas_dengue :(4.00 ;6.00])⟩	0,50	0,853998
⟨(mean_temper :<=23.45 mean_humid :<=76.85 nb_cas_dengue :(4.00 ;6.00]) ⟩	0,50	−0,129980
…	…	…

TABLE 5.18 – MCT pour les données en considérant le zonage *zones d'aménagements*

Motifs	**Support**	**MCT**
⟨(mean_temper :>23.32 ihre_index :<=1.00) (mean_humid :<=77.15)(nb_cas_dengue :<=4.00)⟩	0,80	0,70371
⟨(mean_temper :>23.32 nb_cas_dengue :(4.00 ;5.00])(mean_temper :>23.32)⟩	1,00	−0,142
⟨(mean_temper :>23.32 nb_cas_dengue :(4.00 ;5.00])(mean_wind :<=3.29) (nb_cas_dengue :<=4.00) ⟩	1,00	−0,474
…	…	…

De façon similaire aux résultats expérimentaux sur la pollution, les motifs peu contredits par rapport aux données sont mis en évidence. En effet, même si un motif est plus fréquent qu'un autre, il peut être très contredit par les données (e.g. le troisième motif du Tableau 5.17), donc, moins intéressant pour l'expert. D'ailleurs, nous pouvons constater dans le Tableau 5.18, que les motifs sont très contredits. En effet, cette caractéristique est due à l'absence d'évènements tels que *graveyard*, *community_gather*, *indoor_deposit* entre autres.

En conclusion, notre première méthode basée sur le pré-traitement des données spatialisées a été appliquée aux deux jeux de données réelles et a fournie des résultats intéressants pour les experts. De plus, nous avons montré qu'elle est générique et que le principe peut être adapté à d'autres types de données spatio-temporelles.

5.4 Extraction de motifs spatio-séquentiels

Dans la section précédente, nous avons testé la première méthode proposée dans le Chapitre 3, dans laquelle nous avons regroupé des entités en utilisant certains opérateurs spatiaux. Cette "spatialisation" a pour objectif l'étude de l'impact du groupement des entités spatiales sur l'extraction de motifs séquentiels. Par exemple, dans le jeu de données du bassin versant de la Saône, cette spatialisation a permis de mieux comprendre comment les caractéristiques spatiales des données affectent le processus de fouille de données en prenant en compte deux hypothèses de pollution.

Dans cette section, nous allons extraire un nouveau type de motifs appelé *motifs spatio-séquentiels*. Il est important de rappeler que ce type de motif permet d'étudier l'évolution d'un ensemble de caractéristiques décrivant une zone et son entourage proche au cours du temps. Il est donc essentiel, pour chaque jeu de données, d'avoir une relation de voisinage associant des zones qui partagent une limite en commun dans le cas des quartiers ou qui se trouvent à une certaine distance, comme dans le cas des rivières. Cette relation est matérialisée grâce à une matrice à deux colonnes où les identifiants des entités spatiales seront stockés. Par exemple, pour la Figure 1.2 de la Section 1.3, les zones Z_1 et Z_2 sont voisines, contrairement aux zones Z_2 et Z_3, donc, seules les zones Z_1 et Z_2 seront présentes dans la matrice (cf., Figure 3.11).

Un résumé des caractéristiques des données utilisées dans nos expérimentations est présenté dans le Tableau 5.19.

La suite de cette section se déroulera en deux étapes : dans un premier temps, nous allons appliquer nos propositions sur les deux jeux de données réelles afin d'étudier la sémantique des motifs extraits et dans un deuxième temps, nous allons étudier la performance de nos deux propositions algorithmiques ainsi que les mesures d'élagage proposées dans la Section 3.3.

TABLE 5.19 – Caractéristiques des jeux de données réelles

Jeu de données	Nombre de zones	Nombre de dates par zone (moyenne)	Nombre d'items par date (moyenne)	Granularité temporelle	Granularité spatiale
Saône	223	17	12	journalière	100 km^2
Dengue	12	23	11	hebdomadaire	quartiers

5.4.1 Extraction de motifs spatio-séquentiels sur les données associées à la pollution des rivières

Notre première base de données spatio-temporelles est constituée de relevés d'indicateurs biologiques récupères dans les rivières du bassin versant de la Saône. Les caractéristiques associées à ce jeu de données ont été détaillées dans la Section 5.2.1. Dans ces expérimentations, les stations ont été regroupées en utilisant l'approche de spatialisation appelée ϵ-*voisinage* (cf., Section 5.3.1).

Concernant la relation de voisinage, elle a été construite en prenant en compte la grille formée au moment de la construction de zones homogènes regroupant des stations incluses dans un périmètre de 100 km^2. Deux zones sont voisines si la première se trouve en bas, en haut, à droite ou à gauche de l'autre (voir Figure 5.9). Finalement, nous obtenons 4 voisins par zone (sauf pour les zones aux extrémités de l'espace qui ont deux ou trois voisins).

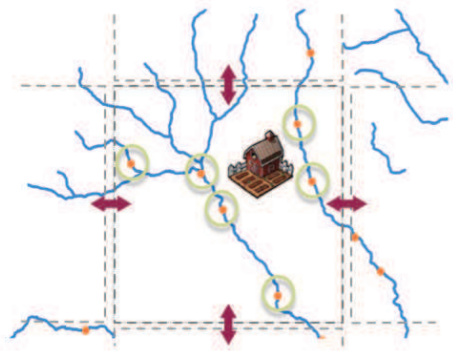


FIGURE 5.9 – Construction de la relation de voisinage pour les stations regroupées en utilisant l'approche de spatialisation ϵ-voisinage

Le Tableau 5.20 montre quelques S2P extraits sur le jeu de données associées au bassin versant de la Saône en utilisant l'algorithme DFS-S2PMiner. Nous pouvons remarquer que nous obtenons une séquence d'itemsets où un itemset est un ensemble d'items (évènements) qui caractérise l'évolution d'un ensemble de stations (couvrant 100 km^2) au cours du temps. Par exemple, la deuxième S2P du Tableau 5.20 ⟨(ibd_ <21)(θ· var_taxo_21-30)(ibd_13-21)⟩ peut être interprétée comme : dans 30% des zones, il existe une diminution de la valeur associée à l'indicateur *ibd*. Cette diminution peut être associée à l'indicateur *variété taxonomique* apparu dans une zone adjacente à la zone étudiée.

La Figure 5.10 illustre la S2P décrite précédemment (cf., ⟨(ibd_<=21)(θ· var_taxo_21-30)(ibd_13-21)⟩). Dans cette figure, nous ne prenons pas en compte la position des stations et la forme des zones pour illustrer les dynamiques spatiale et temporelle du motif. Une éventuelle localisation des zones où ce motif est apparu pourra être effectuée dans l'étape de restitution de motifs.

TABLE 5.20 – Exemples de motifs spatio-séquentiels extraits sur le jeu de données hydrologiques

Motifs	Support
⟨(var_taxo_31-40)(θ·[ibgn_16-20 ; ibgn_etat_bon])⟩	0,32
⟨(ibd_<21)(θ· var_taxo_21-30)(ibd_13-21)⟩	0,30
⟨(ibgn_11-15)(θ·ibgn_etat_bon)(θ·var_taxo_31-40)⟩	0,29
⟨(var_taxo_21-30)(θ·[ibgn_11-15 ; var_taxo_21-30 ; ibgn_etat_moyen])⟩	0,25
…	…

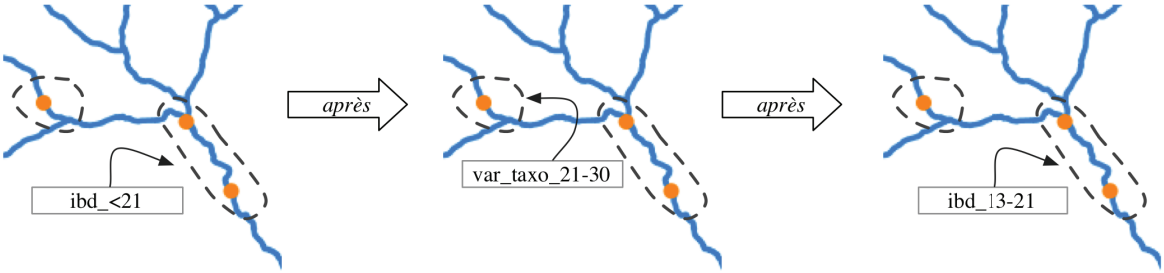


FIGURE 5.10 – Interprétation graphique du motif ⟨(ibd_<=21)(θ· var_taxo_21-30)(ibd :_13-21)⟩

Extraction des S2P peu contredites

Comment discuté auparavant, l'un des inconvénients de la fouille de données est la grande quantité d'informations extraites. Pour réduire le nombre de motifs à explorer pour les experts, nous avons appliqué la mesure de qualité définie dans la Section 4.4 appelée *moindre contradiction spatio-temporelle (MCST)*. Elle a pour objectif, mettre en évidence les motifs les moins contredits par rapport aux données. Le Tableau 5.21 montre les S2P, leur support ainsi que la valeur associée à la moindre contradiction spatio-temporelle.

TABLE 5.21 – Exemples de motifs spatio-séquentiels extraits sur le jeu de données hydrologiques et la valeur associée à la MCST

Motifs	Support	MCST
⟨(θ·var_taxo_21-30)(θ·[ibgn_11-15 ; var_taxo_<21 ; ibgn_etat_moyen])⟩	0,20	0,74501
⟨(ibgn_11-15)(θ·ibgn_etat_moyen)(θ·var_taxo_31-40)⟩	0,33	0,43454
⟨(var_taxo_31-40)(θ·[ibgn_16-20 ; ibgn_etat_bon])⟩	0,32	0,1215
⟨(ibd_<=21)(θ·var_taxo_21-30)(ibd_<21)⟩	0,33	−0,039
…	…	…

Dans le Tableau 5.21 nous constatons que le premier motif, même s'il n'a pas un support très élevé est peu contredit par les données.

5.4.2 Extraction de motifs spatio-séquentiels sur les données associées au suivi épidémiologique de la dengue

La deuxième base de données spatio-temporelles utilisée pour nos expérimentations est la base de données épidémiologiques de suivi de la dengue. Ces données ont été collectées à Nouméa sur un territoire divisé en 12 quartiers. Cette base de données comporte 24 dates pour lesquelles nous disposons d'informations discrétisées décrivant des caractéristiques associées à chaque quartier. Au total, nous avons 11 attributs d'analyse D_A, un attribut associé à la dimension temporelle D_T et un attribut associé à la dimension spatiale D_S incluant les codes des quartiers.

Concernant la relation de voisinage, elle a été proposée pour la mairie de Nouméa. Un quartier est voisin d'un autre s'ils ont une frontière en commun. Par exemple, dans la Figure 5.11 le quartier *Motor Pool* est voisin de *Ngéa Ste Marie*, donc, les identifiants de ces deux quartiers feront partie de la relation de voisinage.

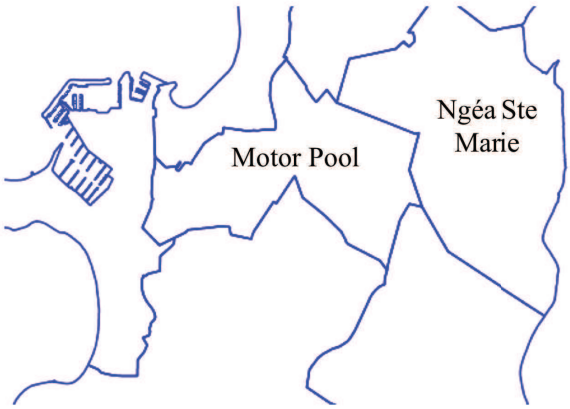


FIGURE 5.11 – Quartiers voisins à Nouméa

Les Tableaux 5.22 montre quelques motifs spatio-séquentiels extraits au cours de l'exécution de l'algorithme DFS-S2PMiner sur les données de surveillance de la dengue. Ces résultats ont été obtenus pour un support minimum fixé à $0,6$.

Lorsque nous étudions les motifs spatio-séquentiels obtenus à partir du jeu de données de surveillance de la dengue, nous pouvons observer qu'ils prennent en compte l'environnement voisin. Par exemple, le troisième motif du Tableau 5.22 peut être interprété par : au moment t_0, nous avons trouvé une présence faible de précipitations et des dépôts privatifs d'eau dans deux zones voisines différentes, puis, dans la zone étudiée, des cas de dengue

TABLE 5.22 – Exemples de motifs spatio-séquentiels extraits sur le jeu de données de la dengue

Motifs	Support
⟨(mean_wind :<=3.20 mean_temper :<=23.55)(θ·[waste_container :<=39.00 ; comunity_gather :<=20.00 ; nb_cas_dengue :<=6.00 ; ihre_index :<=24.55])⟩	0,7
⟨(ihre_index :>34.82 nb_cas_dengue :<=6.00)⟩	0,6
⟨(θ·[precip :<=0.10 ; indoor_deposit :(2126.00 ;2692.50]]) (nb_cas_dengue :<=6.00)(θ·ihre_index :<=24.55)⟩	0,6
⟨(θ·ihre_index :>34.82)(nb_cas_dengue :<=6.00 mean_temper :<=23.55) (θ·comunity_gather :<=20.00)(θ·[nb_cas_dengue :<=6.00 ; ihre_index :<=24.55])⟩	0,6
…	…

sont apparus. Plus tard, nous constatons le développement de nids de moustiques dans une zone voisine.

Extraction des S2P peu contredites

En suivant le même protocole utilisé dans les sections précédentes, nous avons appliqué une mesure objective de validation (MCST) sur les motifs extraits pour filtrer les plus pertinents pour les experts.

Pour les deux jeux de données sur l'eau et la dengue, nous avons montré que nos méthodes permettaient d'obtenir des motifs qui sont sémantiquement pertinents pour les experts.

Nous allons maintenant évaluer le passage à l'échelle de la deuxième approche puisque la première a déjà été démontrée dans la littérature [Chand *et al.*, 2012].

Le Tableau 5.23 montre les S2P, leur support ainsi que la valeur associée à la moindre contradiction temporelle pour le jeu de données associées au suivi épidémiologique de la dengue.

TABLE 5.23 – Exemples de motifs spatio-séquentiels extraits sur le jeu de données de la dengue et la valeur associée à la MCST

Motifs	Support	MCST
⟨(ihre_index :>34.82 nb_cas_dengue :<=6.00)⟩	0,6	0,802595
⟨(mean_wind :<=3.20)(comunity_gather :<=20.00 ·[waste_container :<=39.00 ; ; nb_cas_dengue :<=6.00)⟩	0,76	0,540
⟨(θ· ihre_index :>34.82)(θ·[nb_cas_dengue :<=6.00 ; ihre_index :<=24.55]) (θ·comunity_gather :<=20.00) (nb_cas_dengue :<=6.00 mean_temper :<=23.55)⟩	0,64	0,324551
⟨(θ·[precip :<=0.10 ; indoor_deposit :(2126.00 ;2692.50]]) (nb_cas_dengue :<=6.00)(θ·ihre_index :<=24.55)⟩	0,6	−0,043037
…	…	…

5.4.3 Evaluation de la performance de nos approches

Dans cette section, nous allons utiliser des données réelles ainsi que des données synthétiques décrites dans les Tableaux 5.19 et 5.24 pour évaluer la performance de nos propositions en nous posant cinq questions :

TABLE 5.24 – Caractéristiques des jeux de données synthétiques

Jeu de données	Nombre de zones	Nombre de dates	Nombre d d'items
Graph10x50 (graphe)	10	50	5
Graph10x70 (graphe)	10	70	5
Graph20x70 (graphe)	20	70	5
Graph20x100 (graphe)	20	100	5
Grid20x100 (grille)	20	100	5

1. Quel algorithme est le plus efficace ?
2. Quel est l'impact de la topologie de la relation de voisinage sur le processus d'extraction de motifs spatio-séquentiels ?
3. Quel est l'impact du nombre de zones et du nombre de dates par séquence sur le processus d'extraction ?
4. Par rapport aux mesures d'élagage, l'indice de participation spatio-temporel sera-t-il effectif ?
5. Quel est l'impact de la densité des données sur le processus d'extraction de motifs spatio-séquentiels ?

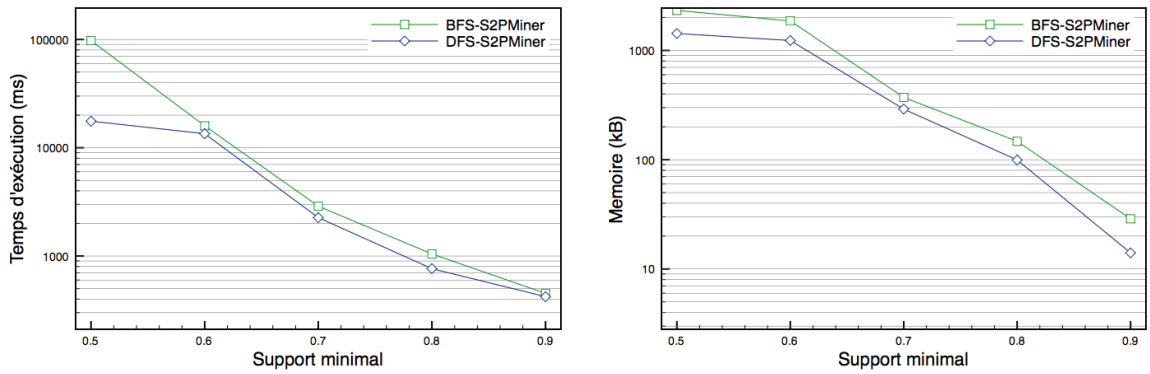
Pour répondre à la première question, nous appliquons nos deux propositions algorithmiques (basées sur les stratégies *Apriori-like* et *pattern-growth*) sur des données synthétiques contenant 10 zones, 70 dates et 5 items.

Pour répondre à la deuxième question, nous étudions l'impact de la topologie de la relation de voisinage en utilisant deux jeux de données synthétiques contenant 10 zones, 50 dates et 5 items - en moyenne - et en prenant en compte deux types de relation de voisinage : (1) la relation de voisinage en forme de grille ; et (2) la relation de voisinage en forme de graphe (cf., Section 5.2.4).

Pour répondre à la troisième question, nous étudions aussi l'impact du nombre de zones sur le processus d'extraction. Pour cela, nous faisons varier le nombre de zones (10 et 20 zones par jeu de données) et nous appliquons l'algorithme *DFS-S2PMiner* sur ces jeux de données. L'impact du nombre de dates par séquence sur le processus d'extraction a également été étudié en faisant varier le nombre de dates par séquence (50 et 70 dates) sur les jeux de données synthétiques.

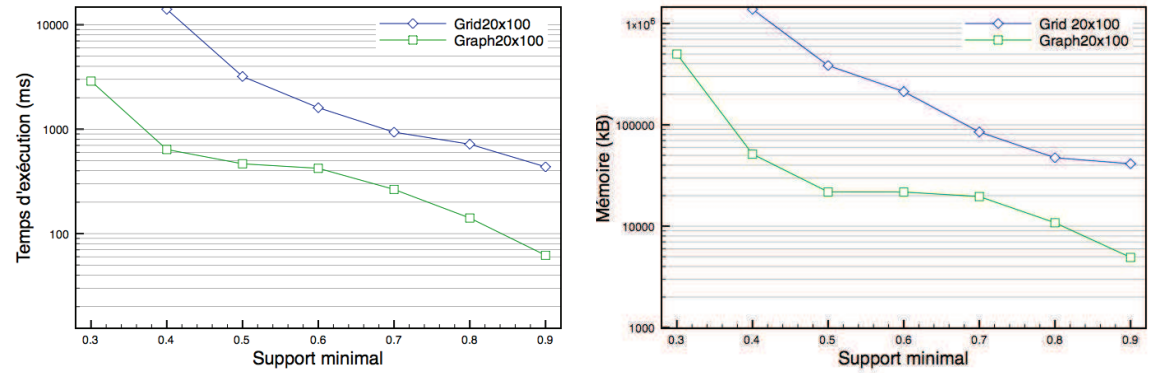
Finalement, pour répondre aux deux dernières questions, nous avons utilisé deux jeux de données réelles pour évaluer l'efficacité de la mesure d'élagage et pour évaluer la performance globale de nos approches.

Les deux algorithmes proposés ont été développés en Java. Les expérimentations ont été réalisées sur un PC basé sur Intel (R) Xeon (R) avec 16 Go de RAM avec Ubuntu Server 9.10 comme système d'exploitation. Les résultats sont discutés dans la section suivante.



(a) (b)

FIGURE 5.12 – Comparaison de l'efficacité des deux algorithmes proposés (BFS-S2PMiner et DFS-S2PMiner) en regardant : (a) le temps d'exécution (b) la mémoire utilisée



(a) (b)

FIGURE 5.13 – Evaluation de l'impact des différents topologies dans le relation de voisinage en considérant : (a) le temps d'exécution (b) la mémoire utilisée

Evaluation quantitative des résultats

Pour évaluer l'efficacité des deux approches (*pattern-growth* ou *Apriori-like*), nous appliquons nos deux propositions algorithmiques sur des jeux de données synthétiques en utilisant relation de voisinage en graphe comportant 20 zones, 70 estampilles temporelles et 4 voisines par zone (Graph20x70). Les Figures 5.12a et 5.12b montrent le temps d'exécution ainsi que la mémoire allouée respectivement pour différents supports. Nous pouvons observer que l'algorithme basé sur une approche en profondeur est plus efficace en temps d'exécution et utilisation de la mémoire. Cette conclusion est soutenue par de nombreux articles dans la littérature (voir par exemple, [Chand *et al.*, 2012]).

Nous étudions l'impact de la topologie de la relation de voisinage sur le processus d'extraction de motifs spatio-séquentiels pour différents supports en utilisant deux jeux de données synthétiques : Graph20x100 et Grid20x100. Les Figures 5.13a et 5.13b montrent l'impact de ces deux types de relation de voisinage (grille ou graphe) sur le temps d'exécution et la mémoire allouée respectivement. Sur ces figures, on constate que le temps d'exécution augmente rapidement pour le jeu de données généré en utilisant de la relation de voisinage basée sur la grille (8 voisins par zone). En revanche, en utilisant la relation de voisinage en graphe, nous obtenons de meilleurs résultats. La relation de voisinage a donc un impact très important dans le processus d'extraction, en particulier sur l'occupation de la mémoire.

Par la suite, on cherche à estimer l'impact du nombre de zones et des estampilles temporelles incluses dans une séquence sur le processus d'extraction en utilisant l'algorithme *DFS-S2PMiner*. La Figure 5.16 montre l'impact de la variation du nombre de zones et d'estampilles temporelles sur le processus d'extraction de motifs spatio-séquentiels en utilisant quatre jeux de données : Graph10x70, Graph20x70, Graph10x50 et Graph10x70. Il est clair que, pour le jeu de données contenant 20 zones et 70 dates par séquence, le temps d'exécution et l'occupation mémoire sont plus élevés. On peut remarquer aussi que lorsque nous augmentons le nombre de dates, les ressources utilisées par l'algorithme augmentent plus que lorsque nous augmentons le nombre de zones.

Enfin, nous évaluons l'efficacité de la mesure d'élagage proposée dans la Section 3.3.2 appelée *l'indice de participation spatio-temporelle* (STPI) et nous montrons les résultats globaux. Volontairement, nous appliquons l'algorithme *DFS-S2PMiner* sur les deux jeux de données réelles.

À cet effet, la Figure 5.17 montre que le nombre de motifs extraits en utilisant la mesure *STPi* est inférieur au nombre de motifs extraits en utilisant le support classique. Ces résultats soulignent que la mesure STPi est très restrictive en raison de l'indice de participation temporelle (TPi). En effet, ce support filtre les résultats en ne conservant que les motifs qui apparaissent à plusieurs reprises dans la base de données spatio-temporelles. Nous voyons ce phénomène dans les jeux de données de suivi épidémiologique de la dengue et plus clairement sur le jeu de données hydrographiques (voir figure 5.17).

D'un autre côté, les résultats montrent la grande efficacité de notre algorithme sur les

données hydrologiques en raison de la faible granularité temporelle et par conséquent, la faible densité des données. Pour ce jeu de données, le processus de fouille commence par un support minimum de $0,5$ et on obtient de bonnes performances, même en utilisant un seuil minimal très bas.

Nous pouvons remarquer que, pour l'ensemble de données denses (e.g. données de surveillance dengue), le temps d'exécution et l'occupation mémoire sont plus élevés. En revanche, pour les jeux de données contenant un nombre important de valeurs nulles, le temps d'exécution et l'occupation mémoire des processus d'extraction S2P sont plus faibles (e.g. pour les données hydrologiques).

DFS-S2PMiner a une bonne performance sur les jeux de données contenant des séquences courtes (e.g. des données hydrologiques). Malheureusement, lorsqu'on extrait de longues séquences (e.g. de jeux de données denses) ou lorsqu'on utilise un seuil minimal très faible, les performances de notre algorithme se dégradent considérablement.

Malgré la diminution du nombre de motifs extraits, le temps d'exécution et la mémoire allouée pour STPi et le support restent équivalents pour des seuils plus élevés, mais diffèrents pour des seuils plus bas. Ce résultat peut être expliqué par le fait que le processus de calcul de la mesure STPi augmente la complexité de l'ensemble du processus d'extraction.

Enfin, l'algorithme basé sur la stratégie de recherche en profondeur (*depth-first-search*) est plus efficace que l'algorithme basé sur l'approche *Apriori* (*level-wise*). Cette différence est due à l'étape coûteuse de génération de séquences candidates, utilisé par les algorithmes de ce dernier type.

Une fois les motifs extraits et filtrés, la question est de savoir comment présenter ces motifs aux experts ? Dans la section suivante, nous proposons une approche de restitution et visualisation de motifs adaptée à nos applications et testée sur les données de suivi épidémiologique de la Dengue en Nouvelle Calédonie.

Des contraintes

Lors de l'étape de fouille de motifs spatio-séquentiels, nous extrayons un nombre très important de motifs. Comme dans la plupart des méthodes de fouille de données, tous les motifs ne sont pas nécessairement intéressants pour l'expert. Pour réduire le nombre de motifs extraits et rendre plus facile l'évaluation des motifs par les experts, nous avons appliqué des contraintes dans le processus de fouille autres que les mesures d'élagages (i.e., le support et l'indice de participation spatio-temporel). Cette utilisation de contraintes est basée sur les notions suivantes : (1) l'écart entre itemsets : spécifie la durée maximale ou minimale entre deux ensembles d'évènements ; et (2) séquence maximale : la plus grande séquence par rapport à l'inclusion. Les Figures 5.14 et 5.15 illustrent ces deux contraintes.

La décision d'ajouter un écart maximal entre les itemsets d'un motif spatio-séquentiel n'est pas arbitraire. Lorsque l'on étudie des phénomènes climatiques ou phénomènes associés aux épidémies, l'écart existant entre des ensembles d'évènements appartenant à

$$\underbrace{(AB)(B \bullet C)(C)}_{\text{GAP=2}}(C)(\theta \bullet D)$$

FIGURE 5.14 – Contrainte d'écart entre motifs spatio-séquentiels

$$(\mathbf{B} \bullet \mathbf{C})(\theta \bullet \mathbf{D})$$
$$(AB)(\mathbf{B} \bullet \mathbf{C})(C)(C)(\theta \bullet \mathbf{D})$$

FIGURE 5.15 – Contrainte d'inclusion de motifs spatio-séquentiels

deux estampilles temporelles consécutives joue un rôle important. Egalement, extraire les sous-séquences (par rapport aux super-séquences) semble pertinent.

5.5 Discussion

Dans le processus d'extraction de connaissances, nous nous sommes concentrés sur l'extraction de motifs spatio-temporels. Nous avons alors proposé deux méthodes nous permettant d'inclure des caractéristiques spatiales dans les motifs obtenus. La première méthode utilise un algorithme d'extraction de motifs séquentiels largement utilisé par la communauté de fouille de données, tandis que la deuxième, utilise une nouvelle méthode permettant l'extraction d'un nouveau type de motif spatio-temporel appelé motif spatio-séquentiel. Ces deux techniques ont été testées sur deux bases de données réelles qui ont été préalablement traitées afin de diviser l'espace en zones homogènes suivant deux hypothèses de pollution, ainsi que sur plusieurs jeux de données synthétiques. Les résultats obtenus pour ces deux approches sont sémantiquement différents.

Le **premier type de motifs** représente l'évolution d'un ensemble de caractéristiques décrivant un ensemble d'entités spatiales regroupées en utilisant différentes proximités spatiales. Il est important de noter que, en appliquant le même algorithme sur la même base de données pour le même support minimum, mais en utilisant des méthodes de division spatiale, on obtient deux ensembles différents de motifs. Cette différence se reflète non seulement dans le nombre de motifs extraits, mais aussi dans leur constitution. Pour savoir quelle approche de spatialisation est la plus intéressante pour les experts, nous avons appliqué une technique de post-traitement (k-means) permettant le regroupement des motifs en clusters homogènes combinée à la mesure statistique appelée *sum of square errors* (SSE).

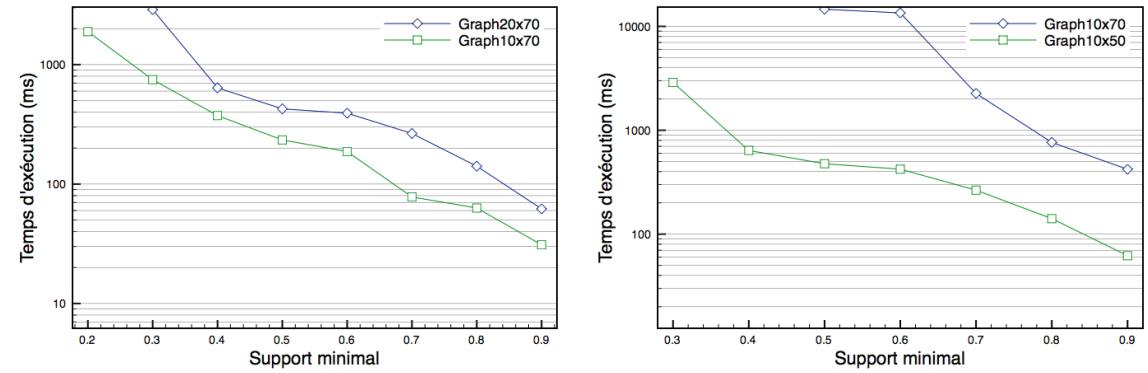
Le **deuxième type de motifs** représente également des changements d'un ensemble d'évènements décrivant des zones au cours du temps. En plus, il contient des informations supplémentaires comme les évènements apparus dans leur voisinage proche. Dans ce type de motif extrait, nous pouvons apercevoir directement les relations spatiales entre les zones voisines grâce à l'opérateur spatial • (à côté de). L'extraction de cette information

supplémentaire affecte directement les performances de nos algorithmes, car l'espace de recherche augmente considérablement avec le nombre de voisins à évaluer. Or cette information supplémentaire peut être cruciale dans les décisions concernant la conservation et la restauration des cours d'eau et leurs milieux environnants, ainsi que pour déclencher une alerte d'épidémie de dengue si une configuration particulière apparaissant dans un motif est reconnue.

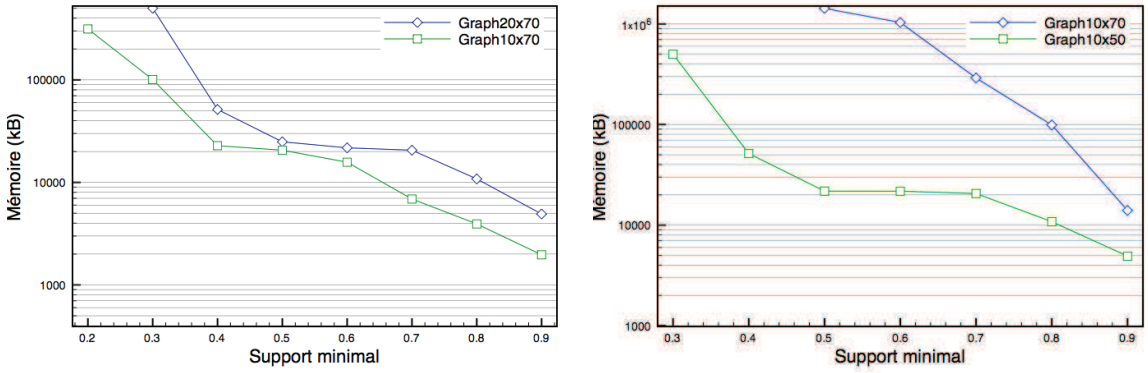
Concernant la granularité temporelle, elle a été prise en compte au moment du pretraitement des données. Effectivement, une maladie comme la dengue est transmise par un vecteur ayant certaines caractéristiques à respecter liées au temps tels que la durée de vie d'un moustique, la période d'incubation des larves, etc. Par exemple, il est connu que les moustiques *Aedes aegypti* ont une période de vie de 10 à 14 jours durant laquelle, les femelles peuvent être potentiellement porteuses du virus de la dengue et par conséquent dangereuses pour les humains.

Dans ce contexte, au moment de l'extraction des motifs, il faut faire attention à ne pas dépasser l'écart entre des itemsets d'une séquence à plus de deux si les données ont une granularité hebdomadaire ou fixer un écart égal à 14 si les données sont journalières. Considérer un écart de temps supérieur à deux semaines n'aura aucun sens quand on extrait de la connaissance sur des données associées à une épidémie. Au contraire, elle pourra avoir du sens si l'on étudie d'autre phénomène comme la pollution des rivières, pour laquelle, les évènements s'étalent sur des longues périodes de temps.

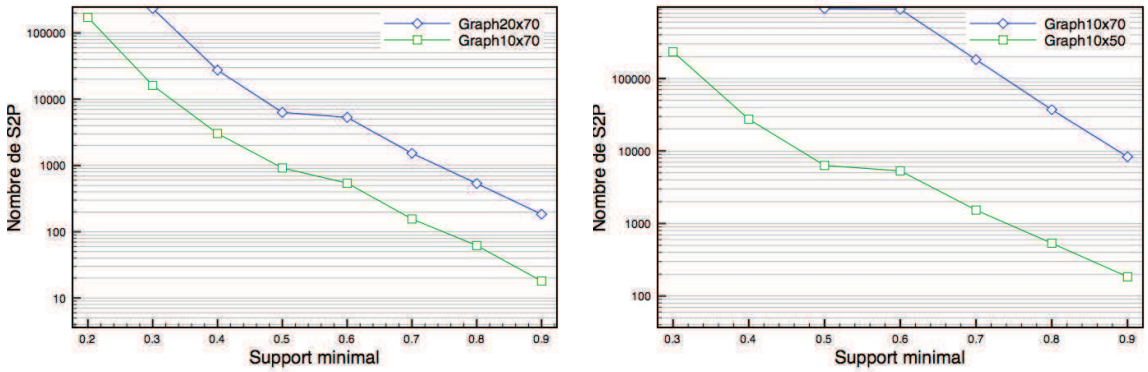
Finalement, les résultats obtenus par Tsoukatos et Gunopulos [2001], bien que proches des notres, ne contiennent pas ces motifs. Comme nous l'avons décrit dans le Chapitre 2, Tsoukatos et Gunopulos [2001] sont à la recherche de séquences d'itemsets qui se produisent fréquemment dans une base de données spatio-temporelles, sans prendre en compte une dynamique spatiale (e.g. le voisinage proche) comme nous l'avons montré dans nos exemples (voir les Tableaux 5.22 et 5.22). Il faut noter que, quantitativement, notre approche et l'approche proposée par Tsoukatos et Gunopulos [2001] ne sont pas comparables en raison de la différence existante entre les espaces de recherche générés pour les deux algorithmes.



(a)

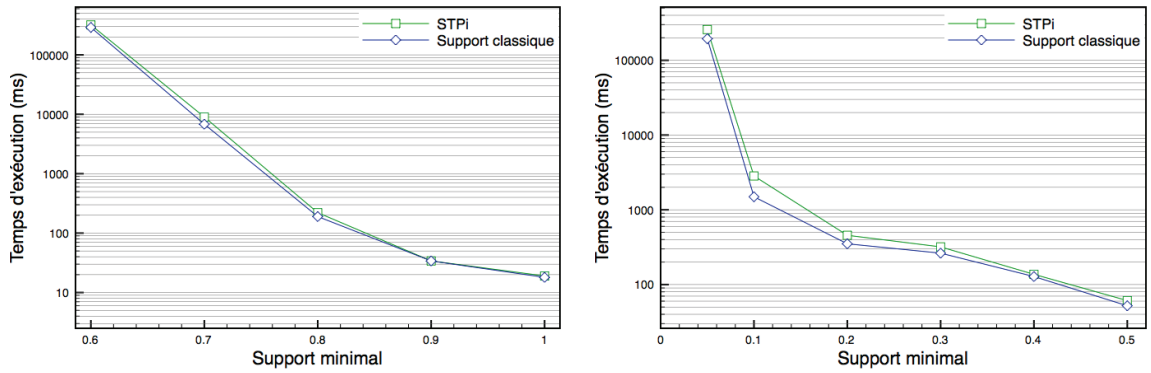


(b)

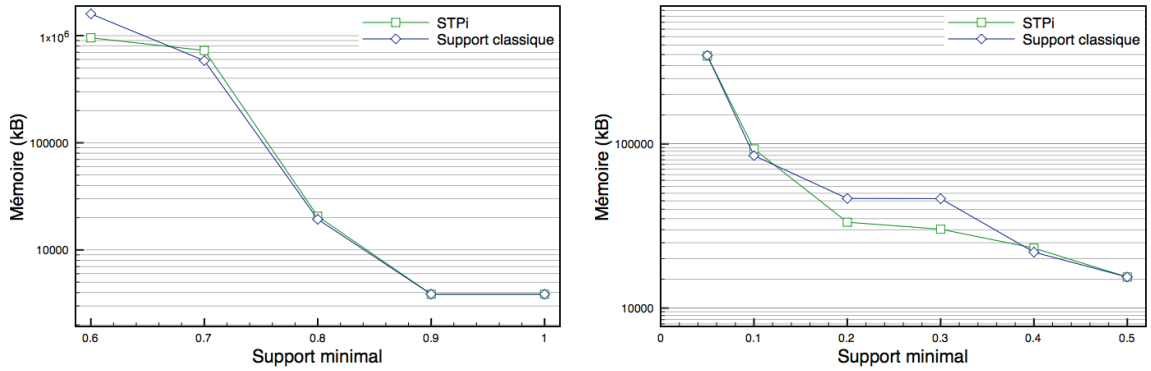


(c)

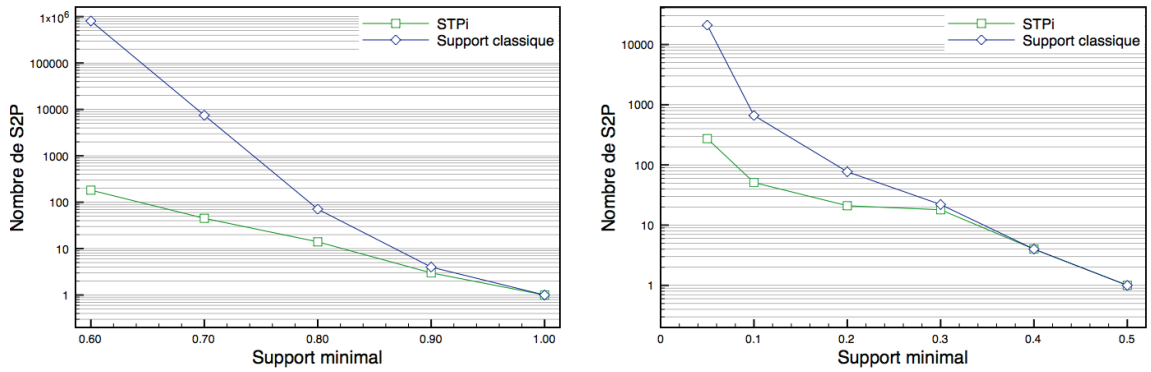
FIGURE 5.16 – Évaluation de l'impact de la variation du nombre de zones et le nombre de dates sur les données synthétiques en prenant en compte : (a) le temps d'exécution (b) la mémoire utilisée (c) le nombre de motifs extraits



(a)



(b)



(c)

FIGURE 5.17 – Evaluation de la efficacité de la mesure d'élagage STPi en utilisant les leu de données de la Saône et la dengue respectivement en regardant : (a) le temps d'exécution (b) la mémoire utilisée (c) le nombre de motifs extraits

Chapitre 6

Visualisation de motifs

Préambule

Les volumes de données collectées et stockées dans des bases de données spatio-temporelles augmentent. Il devient donc crucial de fournir des outils permettant aux experts de mieux appréhender ces données afin de pouvoir prendre des décisions. Toutefois, les motifs extraits sont souvent difficiles à interpréter par les experts. Il est donc indispensable de coupler ces approches à des méthodes de visualisation. La visualisation apporte rapidement une importante valeur ajoutée pour la compréhension de dynamiques spatio-temporelles. Dans ce chapitre, nous présentons une nouvelle approche de visualisation des motifs spatio-temporels.

6.1 Introduction

De nombreux phénomènes évoluent dans l'espace et le temps. La modélisation de ces phénomènes est souvent complexe, non seulement en raison de leur nature spatiale et temporelle, mais aussi à cause des interactions possibles entre les évènements participant aux phénomènes. Les méthodes de visualisation sont alors souvent inadaptées à la complexité de telles données.

Dans notre exemple traitant la météo, l'étude des orages dans une zone en est un exemple typique. Les experts en météorologie savent que la présence d'orages dépend de facteurs environnementaux, (e.g. les rafales, la pression atmosphérique, la température, etc.) ainsi que de la topologie des zones étudiées, (e.g. des montagnes, des plateaux, etc.). Toutefois, l'impact de ces facteurs environnementaux et de ces entités topographiques

reste encore mal connu et il est toujours difficile à prévoir les orages qui sont très localisés. Dans ce contexte, nous avons montré dans ce manuscrit que les méthodes d'extraction de connaissances à partir des données (ECD) apportent des solutions via l'identification sans hypothèse *a priori* de relations entre variables et évènements, caractérisées dans l'espace et dans le temps. Malheureusement, l'exploitation par l'expert des motifs comme ceux obtenus précédemment est souvent limitée. L'expert éprouve souvent beaucoup de difficultés à s'approprier ces nouvelles connaissances qui sont parfois tout aussi complexes à interpréter que les données initiales, notamment lorsque l'on cherche à représenter des dynamiques spatiales et temporelles [Cao *et al.*, 2011]. La mise en place de méthodes et d'outils permettant de mieux restituer ces connaissances est donc un enjeu majeur.

Dans cette section, nous nous focalisons sur la restitution des connaissances obtenues aux experts. Pour cela, nous proposons une approche de visualisation permettant aux experts de mieux appréhender les interactions spatiales et temporelles entre les différents facteurs représentés par les motifs spatio-séquentiels (c.f., Section 3.3). Contrairement aux approches classiques, la méthode de visualisation retenue souligne les dynamiques spatio-temporelles, tout en prenant en compte l'environnement proche. Notre méthode peut être appliquée à d'autres types de motifs, où les dimensions spatiales et/ou temporelles sont présentes, telles que les co-localisations [Shekhar et Huang, 2001] et les séquences temporelles de Tsoukatos et Gunopulos [2001].

Nous ne proposons pas uniquement une méthode de visualisation de motifs spatio-temporels mais tout un environnement permettant de faire une analyse détaillée de ces motifs à différentes échelles (des motifs globaux aux entités spatiales locales). Plus précisément, notre environnement de visualisation offre les avantages suivants :

- un affichage synthétique et schématique des motifs sous forme de graphes colorés (avec possibilité d'associer des icônes aux nœuds). Trois visualisations sont proposées en fonction des besoins des experts et du type de motifs ;
- un affichage détaillé des zones et des dates où sont apparus les évènements (i.e. des occurrences des motifs). À partir d'un motif, il est possible d'identifier les zones impactées sur une carte (et inversement). Une frise chronologique permet de visualiser les dates des évènements représentés par les motifs (avec deux niveaux de détails) ;
- l'affichage de statistiques détaillées sur les zones (e.g. nombre d'habitants) et les caractéristiques temporelles des motifs (e.g. durée moyenne).

À notre connaissance, il n'existe pas d'approche de visualisation qui propose ce type de fonctionnalités, associées à des motifs spatio-temporels aussi complexes. Il est important de noter que notre approche de restitution et de visualisation est générique et peut être appliquée à différents problèmes comme l'érosion des sols, la pollution des rivières, etc. Dans cette thèse, nous avons validé notre approche sur une base de données contenant des informations sur le suivi épidémiologique de la dengue à Nouméa en Nouvelle Calédonie. Les premiers retours des professionnels de santé ont confirmé l'intérêt d'une telle plateforme.

6.2 Travaux préliminaires

La visualisation d'informations est un aspect important pour aider l'expert dans la prise de décisions. Dans ce contexte, certaines techniques visent à représenter visuellement des motifs extraits afin d'aider les experts à mieux comprendre et analyser les informations (voir, par exemple, [Bertini et Lalanne, 2010; Ho *et al.*, 2002]). La visualisation d'informations est le thème central de nombreux articles. Parmi ceux-ci, nous étudions plus en détails ceux mettant l'accent sur la visualisation des résultats obtenus après une fouille de données.

Dans un contexte général, les techniques de visualisation ont été largement discutées dans la littérature (voir par exemple, [Mazza, 2009; Ward *et al.*, 2010]). Ces auteurs insistent, entre autres, sur l'importance de l'affichage visuel des informations de façon à rendre plus facile l'interprétation des résultats obtenus. D'un autre côté, Card *et al.* [1999] conclut qu'un graphisme visuellement attrayant montrant l'information à interpréter, est plus intéressant et souvent plus efficace en terme d'interprétation qu'un affichage immédiat de quelques chiffres ou une représentation purement textuelle [1]. Ces auteurs donnent également des indications sur les spécifications techniques à prendre en compte lors de la représentation visuelle des données. Ils arrivent à la conclusion que la visualisation graphique d'informations doit être précise, claire et efficace. Ces trois caractéristiques doivent être accompagnées de conventions d'ordre technique comprenant : la sélection des couleurs, le choix des formes, les polices, la forme et le remplissage des lignes, le rangement des espaces de design et bien d'autres. Par ailleurs, Keim [2002] propose une classification des techniques de visualisation, selon le type de données à visualiser, e.g. données unidimensionnelles, bidimensionnelles, hiérarchiques, etc.

Plus spécifiquement, les systèmes de visualisation de motifs sont généralement dédiés à des nouvelles techniques d'extraction de connaissances appliquées à différents domaines. En effet, il est très difficile d'aborder le problème de la visualisation des connaissances extraites par la fouille de données, sans aborder la méthode qui précède le système de visualisation. Plusieurs approches ont été proposées pour visualiser des motifs séquentiels. Par exemple, Wong *et al.* [2000] ont appliqué une technique d'extraction de motifs séquentiels aux données textuelles. Cette approche est accompagnée d'un prototype de visualisation permettant l'analyse des motifs séquentiels obtenus sur de grands corpus. Subasic et Berendt [2008] ont proposé une méthode et un outil de visualisation pour cartographier et interagir avec les publications scientifiques postées sur le Web en utilisant des méthodes de fouille de textes. Plus récemment, Sallaberry *et al.* [2011] ont présenté un cadre pour la modélisation et la visualisation de motifs séquentiels permettant d'identifier les associations et les relations hiérarchiques entre des données associées à des puces d'ADN.

Dans le cas des données spatio-temporelles, un état de l'art très complet a été fait par

1. La citation en anglais est *a picture is worth a thousand words.*

Andrienko *et al.* [2003]. Ils présentent un inventaire des techniques d'exploration visuelles existantes en fonction du type de données et des méthodes de fouille de données utilisées. Trois types de données spatio-temporelles ont été étudiés : (1) des données représentant des changements existentiels associés à des entités spatiales, e.g. occurrence ; (2) des données reflétant des changements dans les propriétés spatiales des entités ; et (3) des données représentant des variations temporelles des différents attributs thématiques.

Par ailleurs, Bertini et Lalanne [2010] étudient le rôle de la visualisation et des techniques de fouille de données sur le processus d'extraction de connaissances à partir des données (ECD). Ils distinguent trois catégories de techniques : (1) les techniques fondamentalement visuelles mais qui exigent l'exécution d'un processus de calcul avant la visualisation des résultats ; (2) les techniques où la fouille de données est l'étape prédominante et les résultats sont montrés en s'appuyant sur un système de visualisation ; et (3) les techniques où la fouille de données et la visualisation sont totalement intégrées (il est impossible de distinguer lequel des deux processus joue un rôle prédominant). Ils ont aussi proposé des extensions possibles à ces travaux. Du côté applicatif, Ping *et al.* [2008] ont visualisé des motifs représentant des changements au niveau des entités spatiales) (e.g. expansion de villes). Les motifs extraits étaient des règles d'association spatio-temporelles. Ils les ont représentées en prenant en compte quatre caractéristiques : date, durée, ordre et fréquence. Plus récemment, Burauskaite-Harju *et al.* [2012] ont proposé un ensemble de méthodes permettant, entre autres, la visualisation des dépendances spatiales entre des épisodes des précipitations temporellement synchronisés.

Les travaux étudiant le problème de la visualisation dans le processus d'ECD sont donc nombreux. Toutefois, contrairement à notre approche, ils ne sont pas adaptés à des motifs spatio-temporels complexes faisant à la fois intervenir la dynamique spatiale et l'environnement proche. Un motif tel que *la présence de cas de dengue dans une région est souvent précédée de températures élevées dans une zone située près de réservoirs d'eau* pourrait difficilement être observé via ces méthodes.

6.3 Le prototype S2PViewer

Dans cette section, nous présentons une nouvelle approche de visualisation des motifs spatio-séquentiels. Cette méthode a été intégrée dans un prototype de visualisation appelé S2PViewer. Ce prototype permet de visualiser la dynamique spatiale (voisinage proche) ainsi que la dynamique temporelle. S2PViewer est basé sur Javascript, JQuery, la librairie D3 et Google Earth. Il peut être exécuté sur de multiples plateformes, y compris PC, Mac et Linux. Une version en cours de développement est accessible sur le site http://datamining.univ-nc.nc/ [2].

2. À utiliser préférablement avec la dernière version de Firefox car le plug-in de Google Earth peut ne pas fonctionner correctement avec certaines versions d'autres navigateurs.

6.3.1 Visualisation des motifs

Comme indiqué dans la Section 3.3, un motif spatio-séquentiel (S2P) est une séquence d'itemsets spatiaux. Notre approche de visualisation doit permettre, entre autres, d'avoir un aperçu général de ces motifs, de donner un sens à ce qu'ils représentent et de permettre la découverte d'un motif pertinent parmi les nombreux motifs inintéressants [Bertini et Lalanne, 2010]. Pour représenter visuellement de tels motifs, il faut prendre en compte deux dynamiques : la dynamique spatiale, représentée par les itemsets spatiaux (l'opérateur spatial) et la dynamique temporelle, représentée par l'aspect séquentiel.

Un itemset spatial représente l'état courant d'une zone (ses évènements ou caractéristiques) ainsi que celui de ses voisins proches, à un instant donné. Trois types d'itemsets spatiaux peuvent être représentés à l'aide des opérateurs tels que · (voisin), [] (groupement) et le symbole θ (absence). La Figure 6.1 illustre ces trois cas en utilisant une représentation à base de graphe.

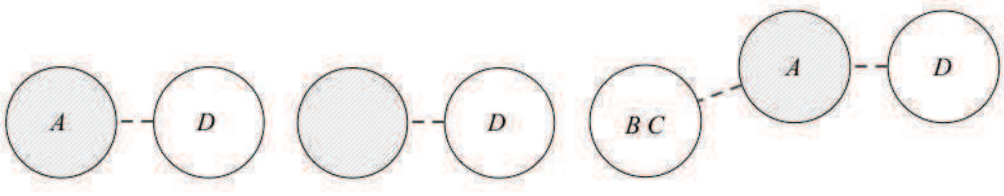


FIGURE 6.1 – Représentation graphique des itemsets spatiaux (a) A·D (b) θ·D (c) A·[BC;D]

Dans la Figure 6.2, chaque cercle représente une zone. Dans les trois figures, la zone colorée est la zone étudiée. Les lignes pointillées représentent le voisinage spatial. Cette représentation du voisinage spatial a été définie par Peuquet [1994] comme une représentation de la *contiguïté spatiale*. La longueur et l'angle des lignes pointillées n'ont pas de signification particulière.

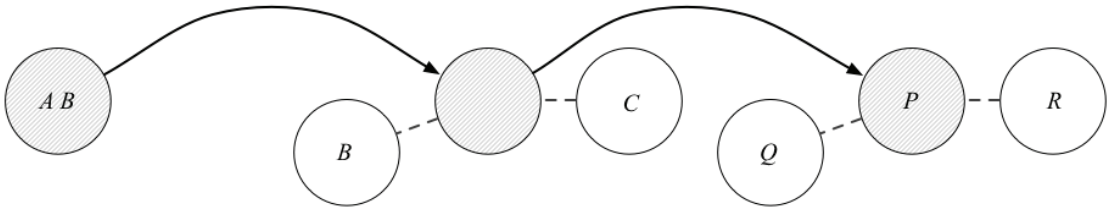


FIGURE 6.2 – Représentation graphique du S2P ⟨(AB)(θ·[B;C])(P·[Q;R])⟩

Les motifs spatio-séquentiels décrivent l'évolution temporelle d'une zone et de son environnement proche à différentes estampilles temporelles. Cette évolution temporelle est

représentée par une succession d'itemsets spatiaux. Nous allons montrer, dans la représentation de ce motif, la dynamique temporelle existant entre itemsets spatiaux contenus dans le motif. La Figure 6.2 montre un S2P et sa dynamique spatio-temporelle. Le S2P illustré est composé de trois estampilles temporelles. La première est composée de deux évènements A et B apparus dans la zone étudiée. Après, il n'existe aucun évènement dans la zone étudiée mais B et C sont apparus dans des zones voisines. Enfin, P est apparu dans la zone étudiée et Q et R sont apparus autour. Les flèches représentent la dynamique temporelle.

Comme on peut le constater, cette représentation respecte le postulat énoncé par Tufte [1983] : *la quantité d'informations (dimensions) affichées ne dépasse pas le nombre de dimensions présentes dans les données*[3]. En effet, deux dimensions sont considérées : la temporalité est représentée par un arc orienté, qui symbolise une succession d'évènements et l'aspect spatial est représenté par des arcs en pointillés.

L'arc orienté représentant la temporalité peut avoir deux configurations possibles : (1) il peut être dynamique dans le plan et être utilisé pour relier deux itemsets qui sont apparus à deux estampilles temporelles consécutives ; et (2) il peut être fixe le long d'un axe et les itemsets peuvent être placés les uns après les autres afin de représenter la séquentialité. L'avantage du premier type de représentation est la visualisation d'un grand nombre d'entités dans un espace réduit. Cependant, la lisibilité des motifs reste faible. Ce type de visualisation est basé sur des algorithmes de force [Kaufmann et Wagner, 2001].

Contrairement à une représentation dynamique, la représentation fixe est facilement lisible mais le nombre d'estampilles temporelles à visualiser reste limité. Dans ce contexte, nous avons privilégié une visualisation en utilisant des coordonnées polaires au lieu des coordonnées cartésiennes [Bertin, 2010] afin d'obtenir un graphique utilisant l'espace alloué proche du carré de façon optimale. L'angle thêta représente la succession ou le temps. Nous avons choisi ce type de représentation pour deux raisons : (1) le ratio de l'arc étant incrémental, de longs motifs pourront être visualisées sous la forme d'une spirale d'Archimède ; et (2) comme cela a été mentionné auparavant, nous représentons deux dimensions. En suivant la maxime de *Tufte* citée plus haut, une représentation en coordonnées polaires est pertinente par rapport à une représentation utilisant des algorithmes de force[4], qui elle projette n dimensions [Kaufmann et Wagner, 2001].

D'autre part, notre approche de visualisation doit représenter des ensembles d'évènements (itemset) à l'aide d'une forme géométrique. Cette forme géométrique permet de montrer le nombre d'évènements contenus dans un itemset en jouant sur sa taille. De plus, l'interface de visualisation doit permettre à l'expert de reconnaître des itemsets contenant un évènement spécifique et donc d'identifier l'apparition d'un évènement ciblé. Un évènement spécifique dit aussi "intéressant" pourra être, par exemple, la présence de dengue et

3. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.

4. *Force-directed graph drawing.* en anglais. L'algorithme de force permet de positionner les nœuds d'un graphe dans l'espace en utilisant un système de force appliqué sur les nœuds et les arcs.

être choisi comme centre d'intérêt par l'expert. Une solution efficace pour représenter les types d'entités (e.g. présence de dengue ou absence de dengue) est d'utiliser des régions fermées et colorées [Ware, 2004]. Nous avons choisi des cercles de deux couleurs différentes : le vert et le rouge car se sont des couleurs opposées dans le modèle appelé *color opponent-process model*. En effet, un œil sain (hormis le daltonisme) les distingue parfaitement, contrairement à d'autres paires de couleurs [Ware, 2004]. Aussi, la couleur rouge est souvent employée, par convention, pour représenter les maladies, le danger, etc., contrairement à la couleur verte qui est souvent utilisée pour représenter des personnes saines, des zones non polluées, etc.

Pour finir, nous utilisons des icônes avec différentes couleurs pour représenter et identifier des évènements. En effet, la représentation des évènements par des icônes avec des couleurs variant selon l'intensité (e.g. haute $\rightarrow$ rouge, basse $\rightarrow$ bleu) semble tout à fait pertinente. Cette idée a été soutenue par des experts en épidémiologie que nous avons rencontré récemment. Ce dernier élément de notre prototype de visualisation est en cours de réalisation.

En intégrant les conventions citées précédemment et en choisissant des disques comme forme géométrique pour représenter les itemsets, des arcs orientés reliant les disques pour la temporalité et des lignes en pointillés pour exprimer la dynamique spatiale, trois types de visualisation ont été proposées dans notre approche :

– *visualisation en graphe* basée sur un algorithme de force, permettant de placer les itemsets dans l'espace et les relier dynamiquement grâce à des arcs orientés. Ce type de visualisation n'est pas adaptée à notre problématique car, comme il a été indiqué précédemment, nous voulons visualiser deux dimensions (cf., Figure 6.3). Cependant, ce type de visualisation reste dans les options de notre prototype pour visualiser des motifs à n dimensions, par exemple, dans le cas de la propagation d'une épidémie ;

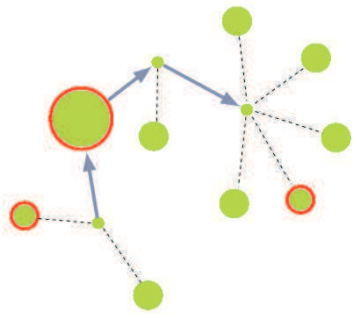


FIGURE 6.3 – Visualisation en utilisant l'algorithme de force

– *visualisation en arc*, permettant de montrer un nombre limité d'itemsets mais est facilement interprétable. Cette approche de visualisation est sélectionnée par défaut ;

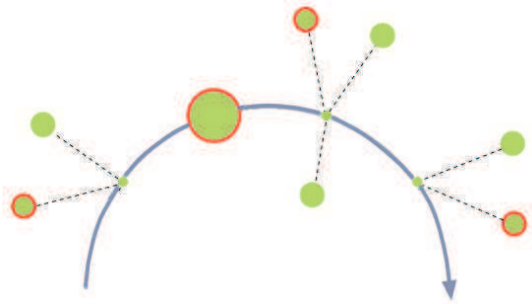


FIGURE 6.4 – Visualisation en arc

– *visualisation en spirale* basée sur la représentation en arc. Ce type de visualisation est très intéressant dans le cas des motifs longs. Effectivement, une spirale d'Archimède favorise la visualisation de séquences comportant de nombreux itemsets à l'aide du rayon incrémental [5]. Toutefois, la perception du déroulement temporel des évènements est difficile à appréhender.

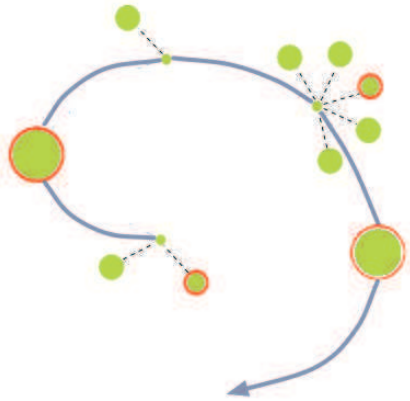


FIGURE 6.5 – Visualisation en spirale

Actuellement, une nouvelle visualisation est sur le point d'être intégrée au S2PViewer. Elle est basée sur la technique *sunburst* [6]. Dans ce nouveau type de représentation, nous conservons l'idée de la visualisation en arc, cependant, les liens spatiaux sont représentés non pas sous la forme d'arcs ou d'arêtes mais sous la forme de zones colorées uniformément. Chaque zone colorée représente un itemset et permet de visualiser "proprement" les évènements de l'itemset (représentés par des disques de différentes tailles, comme dans les cas précédents). Un autre avantage est l'agencement des itemsets sans l'utilisation de

5. Le rayon de la spirale est donné par une fonction continue et monotone de l'angle θ.
6. http://www.cc.gatech.edu/gvu/ii/sunburst/

l'algorithme de force. Il n'y a donc pas de croisement des lignes pointillées avec la ligne en trait plein représentant le temps (comme dans les cas précédents). La Figure 6.6, montre ce type de visualisation.

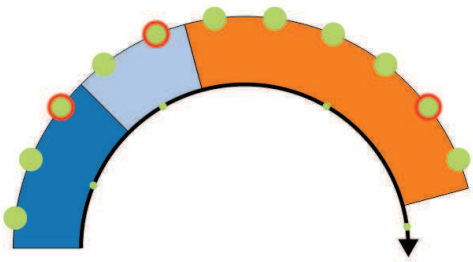


FIGURE 6.6 – Visualisation en utilisant la technique *sunburst*

Nous avons maintenant une idée générale de la visualisation d'un S2P, de la disposition des évènements (itemsets) au cours du temps, tout en prenant en compte les évènements apparus dans des zones voisines. Cette vision générale du motif peut être complétée par une visualisation plus fine du point de vue de la granularité, comme nous le décrivons dans la sous-section suivante.

6.3.2 Vers une visualisation d'informations à différents niveaux

Notre approche de visualisation des motifs S2P se veut globale avec des possibilités de visualisation d'informations détaillées localement sur les éléments d'un motif (e.g. où et quand apparaît ce motif). La visualisation de la dynamique temporelle a pour objectif de valoriser l'information concernant la durée de l'occurrence d'un S2P et le temps où il est apparu. Deux points ont été pris en compte au moment de la représentation de la dynamique temporelle :

- pour chaque motif, une synthèse contenant les périodes, dates du début et de fin de l'apparition du motif dans les zones concernées sera représentée en utilisant des blocs de couleurs différentes. Actuellement, la visualisation des zones impactées pour le motif sélectionné est présentée à l'aide d'une liste. Cependant, une représentation des zones en utilisant une carte est désormais en cours de développement ;
- des statistiques concernant la durée maximale, minimale et moyenne d'un motif, par rapport aux zones où le motif est impacté, seront présentées également par notre approche de visualisation. Ces valeurs calculées "à la volée" permettront à l'expert d'avoir une référence des caractéristiques concernant la durée d'apparition des motifs sur les zones impactées.

Pour présenter la dynamique temporelle (voir Figure 6.7), les conventions à prendre en compte sont différentes de celles considérées pour la représentation de la dynamique

FIGURE 6.7 – Exemple de représentation de la dynamique temporelle d'une séquence spatiale

spatiale. En effet, si l'on utilise des coordonnées polaires, alors une déformation des motifs les plus éloignés sera perçue (ils seront plus longs). De plus, comme cette visualisation n'est pas centrale mais indicative, une représentation en utilisant des coordonnées cartésiennes est privilégiée. Finalement, nous représentons l'apparition d'un motif par un bloc indiquant les dates de début et de fin de ce motif dans une zone.

Par exemple, soit le S2P $s = \langle(A)(B)\rangle$ et la séquence caractérisant une zone $S = \langle(AC)(D)(D)(BC)\rangle$, nous représentons l'apparition de la séquence spatiale s dans S avec un bloc qui commence au temps t0 et finis au temps t3. Si un S2P apparaît de multiples fois dans une même zone, il sera tracé plusieurs fois avec la même couleur. De même, si le motif apparaît dans plus d'une zone, de multiples blocs seront tracés en utilisant des différentes couleurs.

6.4 Le processus d'analyse spatio-temporelle avec S2PViewer

Notre prototype de visualisation est basé sur la technique décrite par Andrienko *et al.* [2003] appelée *interaction de cartes* dans laquelle, des évènements spatio-temporels sont localisés en utilisant une carte. La Figure 6.8 montre le diagramme de flux du processus d'analyse spatio-temporelle des S2P.

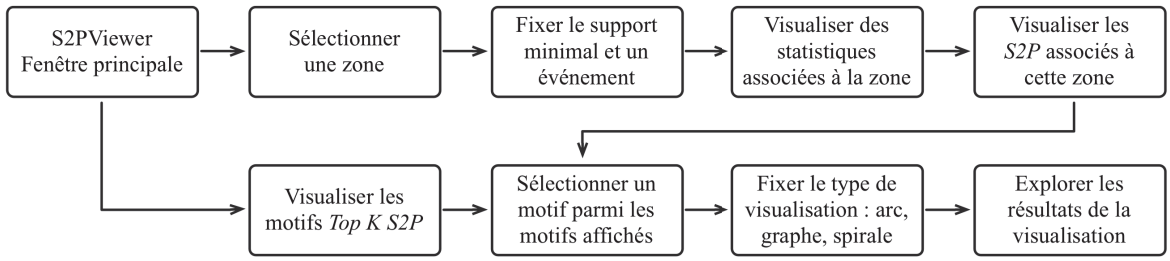


FIGURE 6.8 – Diagramme de flux du processus d'analyse de S2P

Le prototype de visualisation, est composé de trois étapes qui se traduisent par trois vues interactives dans notre interface :

1. La première vue – fenêtre principale – permet de sélectionner une zone. La Figure 6.9 montre l'exemple des données épidémiologiques (données dengue en Nouvelle Calédonie) avec comme zones spatiales les quartiers de Nouméa. Dans cette vue, une seule zone pourra être sélectionnée à chaque fois que l'on désire afficher des motifs la décrivant. Une fois la zone sélectionnée, un support minimal doit être choisi (le support représente la probabilité d'apparition de ces motifs). En fin dans cette vue, l'utilisateur devra sélectionner un évènement considéré comme plus intéressant pour lui, sinon un évènement (par exemple dengue) est sélectionné par défaut (voir Figure 6.12).

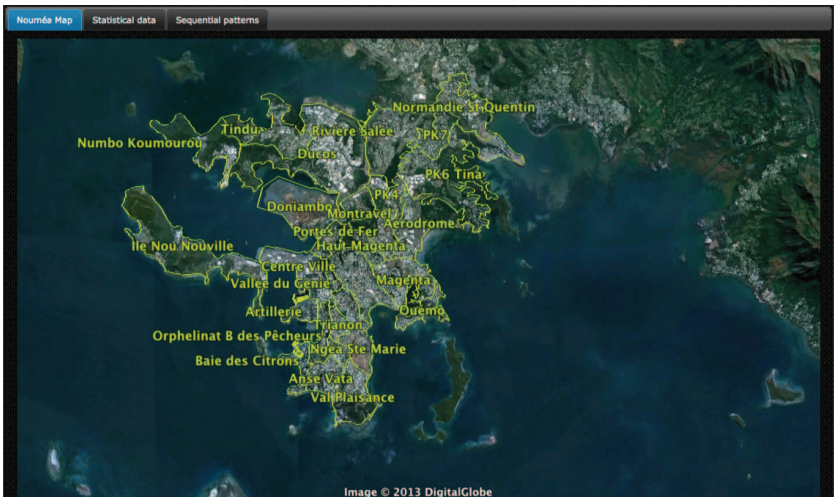


FIGURE 6.9 – Sélection d'un quartier

2. La deuxième vue montre des informations statistiques décrivant la zone choisie dans la vue précédente (voir Figure 6.10). Notamment, nous avons ajouté de l'information concernant la population de la zone (recensement de l'année 2004), le nombre de cas de dengue par année et des informations sur les zones voisines, i.e. ceux qui partagent une frontière commune avec la zone sélectionnée auparavant.
3. Enfin (Figure 6.11), la troisième vue montre les S2P appartenant à la zone sélectionnée ayant un support supérieur ou égal au support fixé dans la première vue. Cette vue interactive permet, la sélection d'un motif, la visualisation graphique de ce motif ainsi que la visualisation de la dynamique temporelle associée (voir Figures 6.13 et 6.14). Si aucune zone n'a été sélectionnée via la première vue, les vingt plus longs motifs seront automatiquement visualisés.

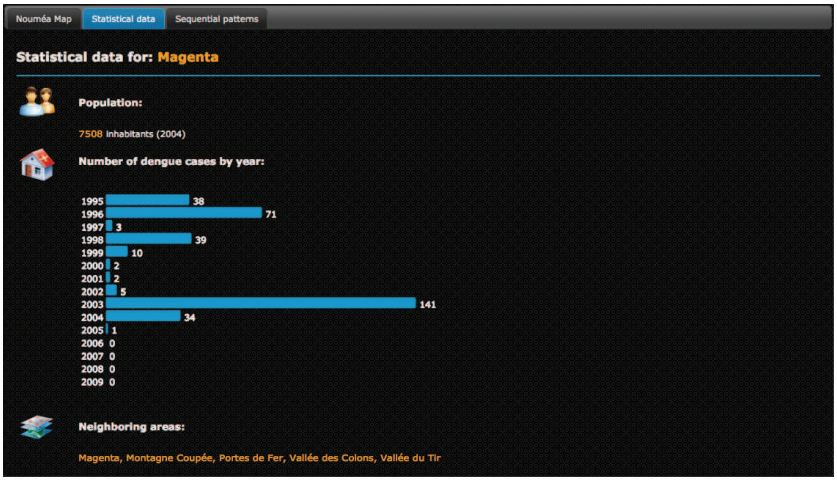


FIGURE 6.10 – Visualisation des informations associées au quartier sélectionné

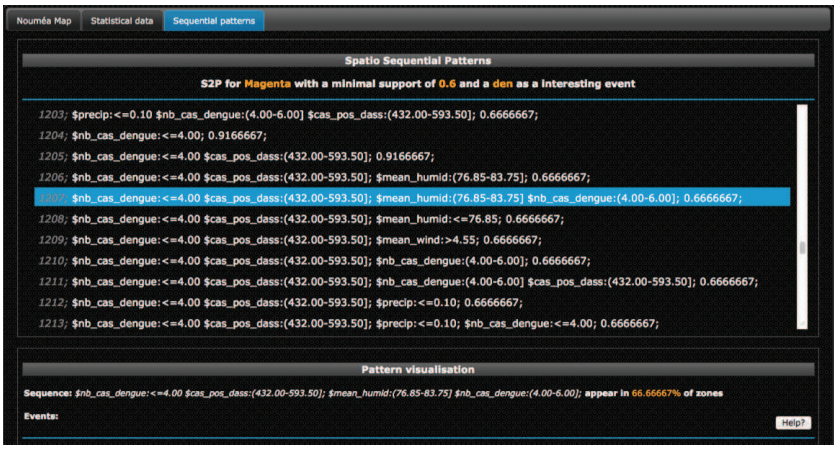


FIGURE 6.11 – Visualisation des 20 plus longs motifs ou des S2P associés au quartier sélectionné

6.5 Analyse sémantique des motifs obtenus

La Figure 6.13 représente visuellement le S2P ⟨(θ·[precip :<=0.10 ; ihre_index :>34.82 ; nb_cas_ dengue :<=6.00]) (mean_wind :<=3.20 nb_cas_dengue :<6.00 mean_temper : <=23.55) (θ·[nb_cas_dengue :<6.00 ; mean_wind :<=3.20 ; mean_temper :<=23.55]) (θ·[mean_humid :(76.85-83.30] ; nb_cas_dengue :<=6.00 ; ihre_index<= 24.55])⟩. Comme nous l'avons décrit dans la Section 6.3.1, chaque cercle symbolise un ensemble d'évènements apparaissant à un moment donné. La taille des cercles reflète le nombre d'évènements caractérisant la zone. L'absence d'évènement dans la zone choisie (i.e. θ) est

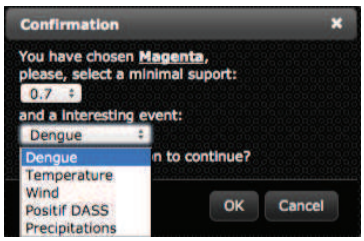


FIGURE 6.12 – Sélection du support minimal et d'un évènement intéressant

représentée par un cercle de petite taille. L'arc de cercle représente le temps et les lignes pointillées représentent la dynamique spatiale (à côté de).

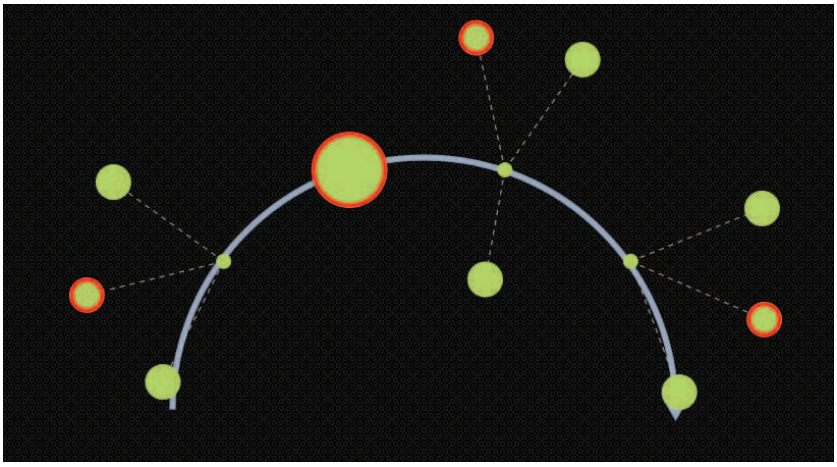


FIGURE 6.13 – Représentation du S2P

La couleur rouge sert à identifier les itemsets qui contiennent l'évènement dengue ou tout autre évènement choisi lors de la première étape (voir Figure 6.12). Cette caractéristique permet d'identifier facilement, d'un côté, la position (i.e., l'estampille temporelle) dans le S2P où se trouve l'évènement intéressant et d'un autre côté, les autres évènements appartenant au même itemset.

Dans la représentation de la dynamique temporelle (voir Figure 6.14), chaque bloc coloré représente la date du début et la date de la fin concernant l'apparition d'un motif dans la séquence représentant une zone (données d'origine). Si le motif apparaît plusieurs fois dans une même zone, il sera représenté par plusieurs blocs de la même couleur. Les couleurs varient pour chaque zone et la ligne temporelle représentée dans l'axe X reste à titre indicatif. L'axe Y contient les zones où le motif est apparu. En plus, le nom de toutes les zones est affiché à gauche et permet d'avoir un aperçu du ratio entre les zones contenant le motif et les zones qui ne le contiennent pas.

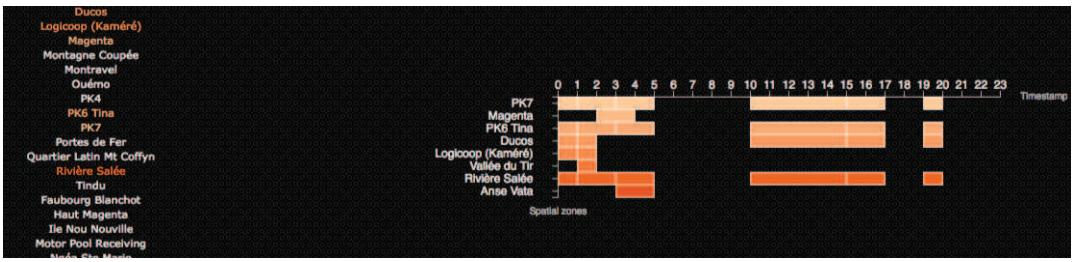


FIGURE 6.14 – Représentation de la dynamique temporelle d'un motif

Comment l'ont mentionné Andrienko *et al.* [2003], il est nécessaire de différencier les évènements temporaires des évènements durables. Les changements de caractéristiques décrivant une zone n'ont pas la même interprétation s'ils apparaissent dans des périodes courtes ou longues. Dans ce contexte, la représentation de la dynamique temporelle de notre approche permet d'identifier aussi bien la durée d'un motif que la périodicité de son apparition dans toutes les zones. Par exemple, la Figure 6.14 illustre le S2P ⟨(nb_cas_dengue :<=6.00) (mean_temper :<=23.55 mean_humid :(76.85-83.30])⟩ qui apparaît périodiquement dans les quartiers *PK7*, *PK6 Tina*, *Ducos* et *Rivière Salée*. Cette information peut devenir cruciale pour les experts quand une épidémie est sur le point de se déclencher.

6.6 Temps de réponse et validation par les experts

Notre prototype de visualisation prend comme fichier d'entrée les motifs obtenus lors de l'étape de fouille de données et transforme cette information en objets à visualiser. Cette transformation ne consomme pas beaucoup de ressources. Au cours de nos expérimentations, nous avons affiché facilement plus de 3 000 motifs en quelques secondes. Ce temps démontre la réactivité de notre prototype. Cette expérimentation met également en avant la nécessité de développer des fonctionnalités pour filtrer et/ou classer les motifs. En effet, il est difficilement envisageable de faire interpréter autant de motifs à un expert. Ce problème a motivé l'intégration dans notre prototype d'une fonctionnalité permettant de visualiser les *top-k* motifs (i.e., les k meilleurs motifs). Dans le prototype, ces motifs sont les plus longs, mais ils pourront être aussi les motifs les moins contredits par rapport aux données.

Concernant la validation de notre prototype, nous avons travaillé en partenariat avec des experts en santé publique de l'Institut Louis Pasteur, de l'IRD et de la Direction des Affaires Sanitaires et Sociales de la Nouvelle Calédonie (DAAS) afin de recueillir des expertises, d'analyser les besoins et ainsi de résoudre les problèmes de lisibilité et d'utilisabilité de notre prototype. Lors de ces collaborations, nous nous sommes intéressés aux données à fouiller et aux besoins des utilisateurs finaux de S2PViewer pour réunir les informations

concernant : (1) l'utilité, i.e. le prototype doit permettre d'aboutir à un résultat et ce résultat doit être pertinent pour ses objectifs ; et (2) l'utilisabilité, i.e. le prototype doit permettre de réaliser une action rapidement et efficacement.

6.7 Discussion

La restitution et la visualisation de motifs est une étape fondamentale du processus d'extraction de connaissances à partir de données. Elle permet, grâce à l'interaction avec les experts du domaine, de transformer les motifs obtenus en connaissances exploitables. Dans cette thèse, nous avons proposé un outil de visualisation de motifs spatio-séquentiels associés au problème de suivi épidémiologique de la dengue en Nouvelle Calédonie. Nous avons décidé de tester notre outil avec les motifs extraits associés aux données de la dengue car nous avions des retours de la part des experts. Cependant, nous pourrons visualiser toute sorte de motifs séquentiels (i.e., séquences d'itemsets), incluant les *motifs spatialement fréquents* extraits grâce à la première approche de fouille présentée dans cette thèse.

S2PViewer a été conçu pour aider les experts à explorer un grand nombre de motifs en visualisant leurs dynamiques spatiale et temporelle. Ce prototype de visualisation permet aussi de montrer les vingt motifs les plus représentatifs de tous les quartiers (*top k*) ou les motifs associés à un quartier préalablement sélectionné. Il permet aussi de connaître certaines caractéristiques associées aux quartiers comme le nombre d'habitants, le nombre de cas de dengue par année et les quartiers voisins à la zone sélectionnée.

Cependant, *S2PViewer* est encore à ses débuts. Actuellement, nous l'améliorons en prenant en compte la représentation condensée d'un ensemble de motifs. Ce type de visualisation permet de représenter graphiquement un ensemble limité de motifs en regroupant les évènements (itemsets) qui sont communs à deux ou plusieurs motifs. Avec cette représentation - qui ressemble à un automate - plusieurs motifs pourront être représentés dans un espace très réduit en utilisant une vue plus abstraite permettant à l'utilisateur de s'orienter dans les données. La Figure 6.15 illustre cette nouvelle approche de visualisation. Dans cette figure, un noeud représente un itemset, il est coloré en rouge si l'itemset contient un évènement intéressant, vert sinon. Chaque séquence est représentée par un ensemble d'arcs de même couleur. Un arc lie un itemset IS_i d'une séquence à un itemset IS_{i+1}. Si n séquences partagent une sous-séquence, il y aura n arcs correspondant à cette sous séquence matérialisés dans cette visualisation. Cette visualisation condensée permet d'identifier des itemsets communs aux séquences et d'optimiser l'espace de visualisation. L'affichage des multiples arcs entre deux itemsets reste à améliorer pour gagner en visibilité.

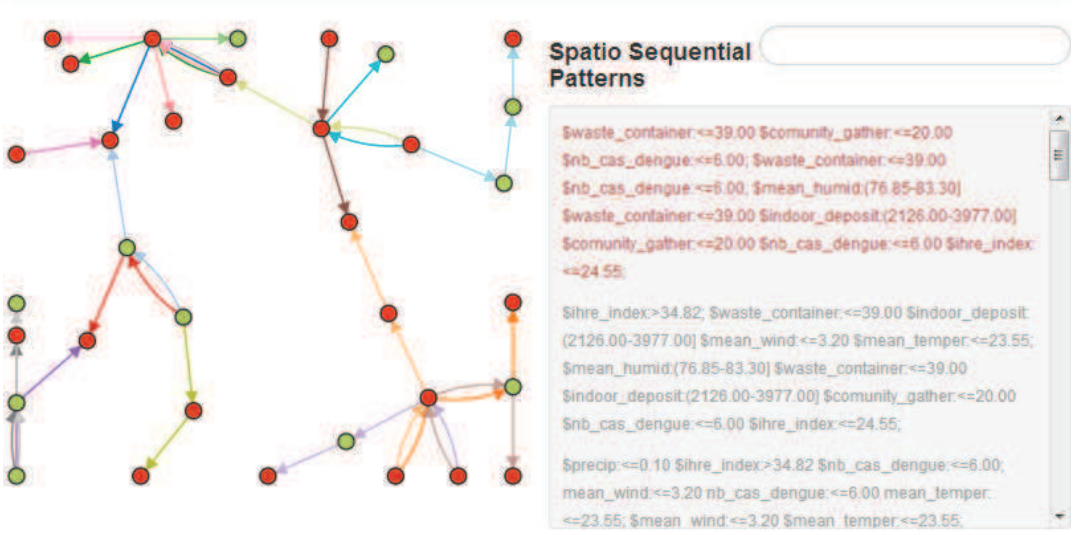


FIGURE 6.15 – Prototype de visualisation de S2P condensés

Conclusions et perspectives

Conclusions

Dans notre vie quotidienne, nous pouvons observer de nombreux phénomènes spatio-temporels. Les évènements météorologiques en sont un exemple. De nouveaux outils, comme les capteurs "intelligents" ayant une forte capacité de stockage, permettent de conserver de grandes quantités de données décrivant ces phénomènes. Cette information est stockée dans des bases de données dites spatio-temporelles. Ces bases comportent des informations associées à un objet spatial, une date d'enregistrement et des informations décrivant l'objet spatial à un moment donné. Par exemple, une interprétation valide d'un enregistrement dans une telle base de données est "le lundi 21 janvier 2013, de fortes pluies ainsi que des orages se sont produits à Paris".

Dans le chapitre dédié à l'état de l'art, nous avons montré qu'il existe de nombreux travaux permettant l'analyse de l'évolution temporelle et spatiale de ces évènements. Pourtant, peu d'approches se focalisent sur les caractéristiques spatiales des objets géographiques, sur les relations existant entre eux et sur leur impact lors de leurs analyses. Par exemple, une ville peut être à côté d'un littoral ou elle peut se trouver dans une région montagneuse qui influe sur sa météorologie.

Dans ce manuscrit, nous avons étudié comment ces relations spatiales impactent le processus d'analyse de données spatio-temporelles selon deux points de vue : (1) la granularité de la relation, qui peut être plus générale ou plus détaillée ; et (2) les relations topologiques, qui définissent comment les objets spatiaux sont disposés les uns par rapport aux autres et dans l'espace en général. Pour cela, nous nous sommes concentrés sur plusieurs étapes du processus d'ECD proposé par Fayyad *et al.* [1996]. Les contributions présentées dans ce manuscrit se résument ainsi :

1. Tout d'abord, un état de l'art sur les approches permettant l'étude des dynamiques temporelles et spatiales des données.

2. Motifs spatialement fréquents : dans cette partie du manuscrit, nous avons utilisé diverses hypothèses permettant de construire des zones homogènes d'objets géographiques. Ce regroupement est réalisé avant l'étape de fouille de données (i.e., dans l'étape de pré-traitement du processus d'ECD). Les zones ainsi construites sont ensuite "fouillées" par un algorithme d'extraction de motifs séquentiels. Les motifs extraits représentent l'évolution d'un ensemble de caractéristiques décrivant une zone au cours du temps. Les expérimentations soulignent l'intérêt de cette proposition en termes de qualité sémantique des résultats. Cependant, pour de nombreux phénomènes spatio-temporels, l'étude de l'évolution des évènements décrivant une zone ne suffit pas car elle ne permet pas de capter la propagation des évènements d'une zone à l'autre.
3. Motifs spatio-séquentiels : dans un deuxième temps, nous proposons un nouveau type de motifs appelé *motifs spatio-séquentiels*. Ce type de motifs permet l'étude d'un ensemble d'évènements décrivant l'évolution d'un objet géographique et de son voisinage proche au cours du temps. Pour cela, nous avons étendu la notion de séquences d'itemsets en ajoutant des opérateurs spatiaux permettant de prendre en compte la notion de voisinage entre entités spatiales. Ensuite, deux mesures d'élagage ont été proposées. Elles sont anti-monotones et permettent de mieux parcourir l'espace de recherche construit lors de l'exécution des algorithmes de fouille. Grace aux motifs spatio-séquentiels, nous pouvons mieux comprendre certains phénomènes spatio-temporels présentant des dynamiques complexes comme dans le cas d'une épidémie et d'une pollution de cours d'eau.
4. Mesures de qualité : le nombre de motifs extraits à partir d'une base de données en utilisant une technique de fouille est souvent très grand et difficile à explorer par les experts. Des motifs dits pertinents peuvent être noyés dans des motifs triviaux. L'intérêt d'utiliser une mesure en post-traitement est justement de faire ressortir ces motifs intéressants. L'une d'elles est la moindre contradiction, qui vise à extraire les motifs les moins ou les plus contradictoires de la base de données. Dans cette thèse, nous avons étendu cette mesure (proposée pour des règles d'association) aux motifs séquentiels et aux motifs spatio-séquentiels.
5. S2PViewer : même si la mesure de filtrage précédente permet de sélectionner les motifs "intéressants", leur exploration reste très difficile. Pour faire face à ce problème, un prototype de visualisation de motifs a été également proposé dans cette thèse. Ce prototype permet de visualiser des motifs séquentiels ainsi que des motifs spatio-séquentiels. Cet environnement propose différents types de visualisation (arc, graphe et spirale) et a été conçu en prenant en compte les remarques des épidémiologistes impliqués dans cette thèse.

Chacune de nos contributions ont été validées par des expérimentations sur des jeux de données réelles associées aux deux domaines d'application que sont la qualité de l'eau

des rivières et le suivi des épidémies de dengue, ainsi que sur des données synthétiques pour tester les limites algorithmiques de nos approches.

Même si dans ce manuscrit, nous avons abordé et résolu différents problèmes associés à l'extraction de connaissances sur des bases de donnés spatio-temporelles, de nombreuses perspectives ont été identifiées et sont décrites ci-dessous.

Perspectives

Nous décrivons dans cette section des perspectives à moyen et long terme associées à nos contributions. Elles concernent les quatre étapes du processus d'ECD que nous avons traitées dans notre manuscrit.

Pré-traitement

Dans cette étape du processus d'ECD, nous avons étudié l'impact des relations spatiales existantes entre des objets géographiques. Ces relations ont été construites afin de favoriser la compréhension de certains phénomènes, comme la pollution des rivières ou l'impact des activités humaines sur la propagation d'une épidémie. Toutefois, un grand nombre de possibilités reste à explorer, par exemple, dans le cas des entités spatiales vagues, i.e., dont les contours ne sont pas complètement délimités. Ce problème entraîne un décalage entre la réalité spatiale et sa description connue. Une solution viable est la modélisation par la logique floue ou en utilisant des modèles probabilistes (notion d'incertitude). Ces perspectives sont tout-à-fait envisageables car dans ce manuscrit, la nature des objets géographiques est clairement définie.

Fouille de motifs spatio-séquentiels

Dans cette étape importante, nous avons identifié plusieurs perspectives et améliorations à apporter. Tout d'abord, concernant la relation de voisinage, elle est actuellement définie par deux objets qui se trouvent l'un à côté de l'autre ou qui partagent une frontière. Deux questions se soulèvent autour de cette définition : (1) existe-t-il d'autres relations de voisinage exploitables ? Par exemple, si l'on souhaite appliquer notre méthode aux données issues de puces à ADN ou sur les images en trois dimensions, une étude approfondie devra être faite afin de bien prendre en compte les possibles relations de voisinage existantes dans ces deux cas ; et (2) pourrait-on utiliser cette notion de voisinage pour étudier une éventuelle propagation du phénomène de zones en zones ? Nous répondons par l'affirmative à cette question. Il suffit d'étendre la notion de voisinage du niveau 1 au niveau n, c'est-à-dire, le voisin du voisin et ainsi de suite. Cette extension de la relation de voisinage peut avoir un impact important sur la performance de l'algorithme associé car l'espace de recherche va augmenter considérablement.

Concernant les résultats obtenus, nous envisageons de réduire le nombre de motifs à explorer en utilisant une représentation condensée des motifs. Celle-ci permettra de regrouper les motifs ayant des itemsets spatiaux communs à l'aide d'un opérateur de groupement des itemsets spatiaux. La question est de savoir, comment intégrer cette représentation à nos algorithmes de fouille tout en conservant leur performance. Une autre solution possible, est l'utilisation de la notion de motifs fermés. Un motif fermé est tel qu'aucun de ses super-motifs n'a un support identique. Peut-on étendre la notion de motif fermé aux motifs spatio-séquentiels ? et comment prendre en compte les informations spatiales et temporelles associées ?

Finalement, nous voudrions améliorer la performance de nos propositions en utilisant le calcul parallèle. En effet, cette technique consiste à exécuter simultanément plusieurs programmes, qui coopèrent pour aboutir à un objectif commun et/ou qui sont en compétition pour la possession de ressources de l'ordinateur. En prenant cette définition, il est envisageable d'utiliser le calcul parallèle dans le processus de fouille de façon à améliorer considérablement les temps de réponses et l'occupation de mémoire de nos algorithmes.

Mesures de qualité

Des questions récurrentes concernant les motifs contenant des informations "complémentaires" (e.g. basse température/haute température) ont souligné le besoin d'extraire des motifs discriminants d'une classe par rapport à une autre. C'est justement l'idée de l'extraction de motifs émergents (emerging pattern en anglais). Dong et Li [1999] définissent un motif émergent comme des ensembles d'évènements qui apparaissent très fréquemment dans une classe et faiblement dans une autre. Extraire ces motifs permet de capturer des changements et des différences significatives entre deux classes d'un même jeu de données.

Dans ce contexte, la moindre contradiction soit temporelle soit spatio-temporelle, peut être étendue et utilisée comme une mesure discriminante. En effet, vu que la moindre contradiction a pour objectif de mettre en évidence les motifs les moins (respectivement les plus) contredits par rapport à la base de motifs contenant un/des évènement(s) que l'on désire comparer, elle peut être utilisée comme mesure discriminante.

Nous proposons, par exemple, à partir de deux bases de motifs $\mathfrak{m}BD_1$ et $\mathfrak{m}BD_2$ extraits à partir de deux jeux de données représentant deux états opposés du même phénomène et un motif p appartenant à $\mathfrak{m}BD_1$. Si p est fortement contradictoire par rapport à la base de motifs $\mathfrak{m}BD_2$, alors, on pourra conclure que p est un motif soulignant une différence existant entre la base $\mathfrak{m}BD_1$ et $\mathfrak{m}BD_2$. En d'autres termes, plus grande est la valeur de la moindre contradiction de p, plus le motif est discriminant par rapport à la base de motifs $\mathfrak{m}BD_2$. En résumé, il s'agit de chercher les motifs appartenant à $\mathfrak{m}BD_1$ les plus contradictoires par rapport à $\mathfrak{m}BD_2$. Cette idée va être bientôt validée et testée sur une vaste gamme de jeux de données pour vérifier son efficacité.

Prototype de visualisation

Les perspectives associées à S2PViewer sont nombreuses. À terme, nous utiliserons des icônes avec différentes couleurs pour représenter et identifier des évènements. En effet, la représentation des évènements par des icônes avec des couleurs variant selon l'intensité (e.g. haute $\rightarrow$ rouge, basse $\rightarrow$ bleu) semble tout à fait pertinente. Cette idée a été soutenue par les experts en épidémiologie impliqués dans le projet.

Nous voudrions aussi réduire le nombre de motifs à visualiser. Pour cela, nous sommes à la recherche de méthodes permettant, par exemple, de montrer les plus longues séquences ou celles contentant un évènement spécifique à une position donnée (e.g. l'apparition de la dengue dans le dernier itemset). Par ailleurs, des améliorations sur la visualisation de motifs sont en cours de développement, notamment la représentation des motifs en utilisant une vue plus abstraite permettant à l'utilisateur de s'orienter dans les données.

Cette liste de perspectives pourrait être étendue car de nombreuses autres voies, aussi bien théoriques qu'applicatives, restent à explorer.

Pour conclure ce manuscrit, lorsque l'on analyse des données spatio-séquentielles, comme pour les méthodes traditionnelles de type clustering, classification, recherche de motifs comme les règles d'association spatiales, les co-locations, les trajectoires, etc. on retrouve des questions assez classiques portant : (1) sur le passage à l'échelle lorsqu'on travaille avec d'importants volumes de données ; (2) sur la qualité des motifs obtenus qui doivent être réellement adaptés aux besoins applicatifs des experts des données ; (3) sur l'intégration de connaissances du domaine (méta-données, ontologies) pour sélectionner, enrichir et valider les motifs.

Toutefois, la nature des données spatio-temporelles impacte également le processus de fouille et des challenges spécifiques ont été identifiés. Tout d'abord, la sémantique des motifs extraits doit être considérée pour présenter aux experts des motifs réellement adaptés à leurs besoins applicatifs. Dans cette thèse, nous avons présenté les motifs spatio-séquentiels qui sont un type de motifs basés sur la notion de séquence. Des motifs à structure plus complexe, comme les graphes, peuvent s'avérer réellement performants dans les bases de données spatio-temporelles comme le montrent les travaux prometteurs de [Sanhes *et al.*, 2013; Pasquier *et al.*]. Par ailleurs, les méthodes de fouille de données spatio-temporelles génèrent souvent de très nombreux motifs, parfois plus nombreux que les données initiales. Il est donc important d'appliquer des contraintes et de définir des mesures d'intérêt comme la moindre contradiction spatio-temporelle, qui permettent aux experts de sélectionner les motifs les plus pertinents. De plus, la définition de visualisations performantes adaptées à ces nouveaux motifs facilitera leur interprétation. Ces visualisations doivent permettre de comparer les motifs et doivent pour cela inclure des outils pour les grouper, pour choisir des représentants à ces groupes ou leur associer des moyennes. Ces outils utiliseront des mesures comme l'inclusion spatiale et/ou temporelle.

Concernant les domaines d'application, beaucoup restent à explorer comme par

exemple la fouille d'images où beaucoup de données existent mais peu de méthodes efficaces et passant à l'échelle ont été développées. Finalement, un réel besoin de collaboration entre experts de la fouille de données et des experts des domaines applicatifs est nécessaire pour évaluer les méthodes et les motifs obtenus.

Annexe A

Annexe

Quartiers de Nouméa, Nouvelle Calédonie (cf., Figure A.1). Source : http://www.sign.nc/presentation/presentation-du-territoire/noumea

FIGURE A.1 – Quartiers de Nouméa

Zones d'aménagement à Nouméa, Nouvelle Calédonie. (cf., Figure A.2) Source : http://www.ac-noumea.nc/histoire-geo/spip/spip.php\?article117

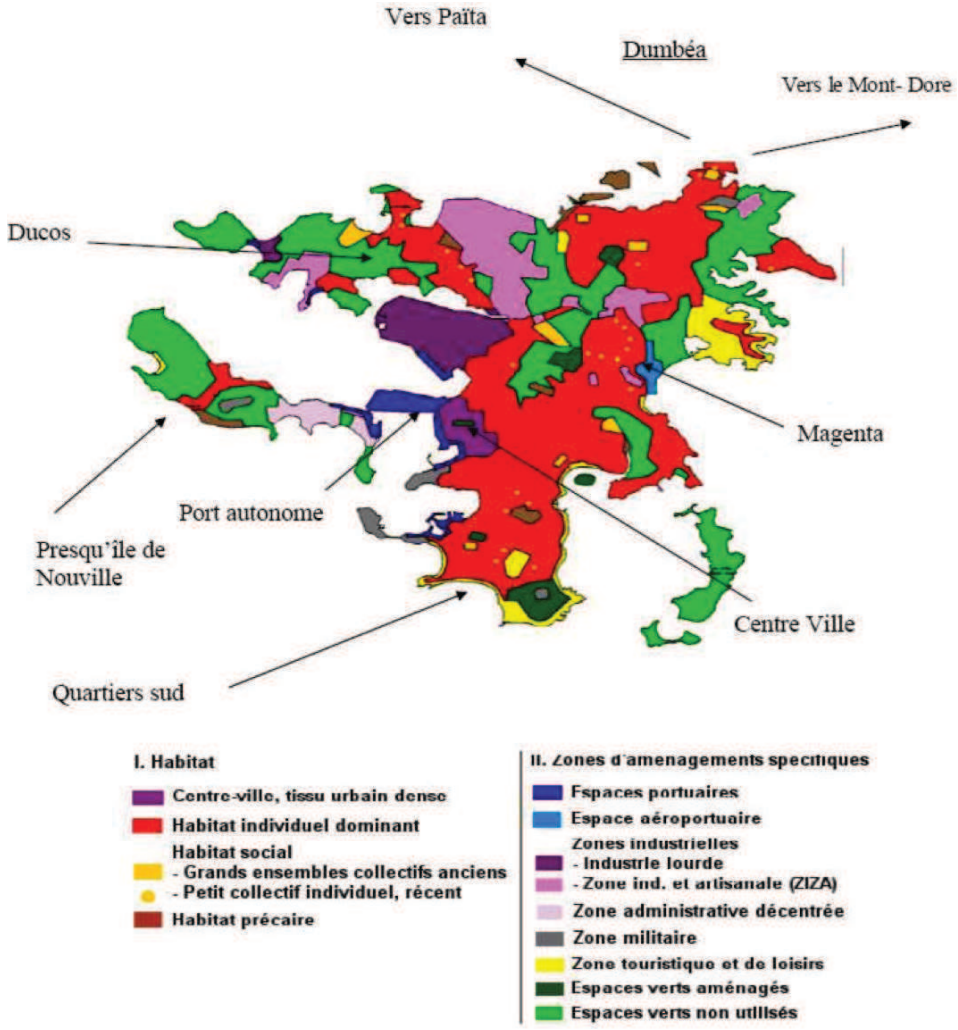


FIGURE A.2 – Zones d'aménagement à Nouméa

Bibliographie

P. Adriaans : *Data Mining*. Pearson Education, 1996. ISBN 9788-13170-717-3. Cité page 33.

R. Agrawal, T. Imieliński et A. Swami : Mining association rules between sets of items in large databases. *In International Conference on Management of data*, SIGMOD '93, pages 207–216, New York, NY, USA, 1993. ACM. ISBN 0-89791-592-5. Cité pages 4 et 60.

R. Agrawal et R. Srikant : Fast algorithms for mining association rules in large databases. *In 20th International Conference on Very Large Data Bases*, VLDB'94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1-55860-153-8. Cité pages 48 et 49.

R. Agrawal et R. Srikant : Mining sequential patterns. *In Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14. IEEE Computer Society, 1995. Cité pages 4, 32, 39, 44, 45, 61 et 80.

N. Andrienko, G. Andrienko et P. Gatalsky : Exploratory spatio-temporal visualization : an analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541, 2003. ISSN 1045-926X. Cité pages 108, 114 et 118.

V. Asproth, A. Hakansson et P. Rèvay : Dynamic information in GIS systems. *Computers, Environment and Urban Systems*, 19(2):107–115, 1995. ISSN 0198-971-5. Cité page 5.

J. Azé : Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. *In Revue RIA-ECA numéro spécial (EGC'03)*, volume 17, pages 171–182, 2003. Cité pages 61, 63 et 65.

J. Azé, P. Lenca, S. Lallich et B. Vaillant : A study of the robustness of association rules. *In The 2007 International Conference on Data Mining*, DMIN'07, pages 132–137, 2007. Cité page 63.

B. Baesens, S. Viaene et J. Vanthienen : Post-processing of association rules. *Departement Toegepaste Economische Wetenchappen (DTEW) Research Report*, 0020:1–18, 2000. Cité pages 59 et 60.

P.E. Bergeret, F. Flouvat et J.M. Petit : Vers la génération de jeux de données synthétiques réalistes pour les motifs fréquents. *In Bases de Données Avancées*, BDA'07, 2007. Cité page 71.

J. Bertin : *Semiology of Graphics : Diagrams, Networks, Maps*. Economic & Social Research Institute, 2010. ISBN 9-781-58948-261-6. Cité page 110.

E. Bertini et D. Lalanne : Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining Explor. Newsl.*, 11(2):9–18, 2010. ISSN 1931-0145. Cité pages 107, 108 et 109.

V. Bogorny, B. Kuijpers et L.O. Alvares : Reducing uninteresting spatial association rules in geographic databases using background knowledge : a summary of results. *International Journal of Geographical Information Science*, 22(4):361–386, 2008. Cité page 61.

V. Bogorny, Engel. P. et L.O. Alvares : Spatial data preparation for knowledge discovery. *In 1th. IFIP Academy On The State Of Software Theory And Practice - PhD Colloquium, Porto Alegre, Brazil*, volume 24, page 8, 2005. Cité pages 30 et 34.

V. Bogorny, J.F. Valiati, S.S. Camargo, P.M. Engel, B. Kuijpers et L.O Alvares : Mining maximal generalized frequent geographic patterns with knowledge constraints. *In IEEE International Conference on Data Mining ICDM*, pages 813–817. IEEE Computer Society, 2006. Cité pages 18, 19, 20 et 28.

W. Boulila, I.R. Farah, K. Saheb Ettabaa, B. Solaiman et H.B. Ghézala : Spatio-temporal modeling for knowledge discovery in satellite image databases. *In Conférence en Recherche d'Information et Applications*, CORIA'10, pages 35–49. Centre de Publication Universitaire, 2010. Cité page 28.

I. Bruha et A. Famili : Postprocessing in machine learning and data mining. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining Explor. Newsl.*, 2(2):110–114, 2000. ISSN 1931-0145. Cité page 60.

A. Burauskaite-Harju, A. Grimvall, C. Achberger, A. Walther et D. Chen : Characterising and visualizing spatio-temporal patterns in hourly precipitation records. *Theoretical and Applied Climatology*, 109(3-4):333–343, 2012. ISSN 0177-798X. Cité page 108.

H. Cao, N. Mamoulis et D. W. Cheung : Discovery of Periodic Patterns in Spatiotemporal Sequences. *IEEE Transactions on Knowledge and Data Engineering TKDE*, 19(4):453–467, 2007. ISSN 1041-4347. Cité pages 16 et 28.

H. Cao, N. Mamoulis et D.W. Cheung : Mining frequent spatio-temporal sequential patterns. *IEEE International Conference on Data Mining ICDM*, pages 82–89, 2005. ISSN 1550-4786. Cité pages 16 et 28.

L. Cao, H. Zhang, Y. Zhao, D. Luo et C. Zhang : Combined mining : Discovering informative knowledge in complex data. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, 41(3):699 –712, 2011. ISSN 1083-4419. Cité page 106.

S.K. Card, J.D. Mackinlay et B. Shneiderman : *Readings in information visualization : using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999. ISBN 1-55860-533-9. Cité page 107.

M. Celik, J. M. Kang et S. Shekhar : Zonal Co-location Pattern Discovery with Dynamic ParaM.rs. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 433–438, octobre 2007. Cité pages 19 et 28.

M. Celik, S. Shekhar, J. Rogers et J. Shine : Sustained Emerging Spatio-Temporal Co-occurrence Pattern Mining : A Summary of Results. *18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, pages 106–115, 2006. ISSN 1082-3409. Cité pages 20 et 28.

M. Celik, S. Shekhar, J.P. Rogers et J.A. Shine : Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE Transactions on Knowledge and Data Engineering TKDE*, 20(10): pages 1322–1335, 2008. ISSN 1041-4347. Cité pages 19, 20 et 28.

C. Chand, A. Thakkar et A. Ganatra : Sequential Pattern Mining : Survey and Current Research Challenges. *International Journal of Soft Computing and Engineering*, 2(1):185–193, 2012. Cité pages 96 et 99.

N. Chelghoum et K. Zeitouni : Spatial data mining implementation : Alternatives and performances. *In Brazilian Symposium on GeoInformatics (GeoInfo'04)*, pages 127–153, 2004. Cité pages 18 et 28.

A. N. Clark et A.N. Clark : *The Penguin dictionary of geography / Audrey N. Clark*. Penguin, London :, 1993. ISBN 0140512330. Cité page 29.

N.A.C. Cressie : *Statistics for Spatial Data (Wiley Series in Probability and Statistics)*. Wiley-Interscience, Revised édition, 1993. ISBN 0-47100-255-0. Cité page 28.

J.M. Diamond : *Armas, Gérmenes y Acero : La Sociedad Humana y Sus Destinos*. Debate pensamiento. Debate, 1998. ISBN 9788483061145. Cité page 87.

G. Dong et J. Li : Efficient mining of emerging patterns : discovering trends and differences. *In Fifth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD'99, pages 43–52, New York, NY, USA, 1999. ACM. ISBN 1-58113-143-7. Cité page 124.

B. Elias : Extracting landmarks with data mining methods. *In Spatial Information Theory. Foundations of Geographic Information Science*, volume 2825, pages 375–389. Springer Berlin / Heidelberg, 2003. ISBN 978-3-540-20148-9. Cité page 33.

J. Eno et C.W. Thompson : Generating synthetic data to match data mining patterns. *IEEE Internet Computing*, 12(3):78–82, 2008. ISSN 1089-7801. Cité page 78.

U.M. Fayyad, G. Piatetsky-Shapiro et P. Smyth : Advances in knowledge discovery and data mining. chapitre Data Mining to Knowledge Discovery : an Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. ISBN 0-262-56097-6. Cité pages vii, 3, 4 et 121.

P. Fisher, P. Laube, M. Kreveld et S. Imfeld : Finding REMO - Detecting Relative Motion Patterns in Geospatial Lifelines. *In Developments in Spatial Data Handling*, pages 201–215. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-26772-0. Cité pages 17 et 28.

F. Flouvat, N. Selmaoui-Folcher, D. Gay, I. Rouet et C. Grison : Constrained colocation mining : application to soil erosion characterization. *In 25th Symposium On Applied Computing*, SAC'10, pages 1054–1059, 2010. Cité pages 19 et 28.

A.A. Freitas : On objective measures of rule surprisingness. *In Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, PKDD'98, pages 1–9, London, UK, UK, 1998. Springer-Verlag. ISBN 3-540-65068-7. Cité page 60.

L. Geng et H.J. Hamilton : Interestingness measures for data mining : A survey. *ACM Computing Surveys (CSUR)*, 38(3), 2006. ISSN 0360-0300. Cité pages 44 et 60.

F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo et R. Trasarti : Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The Very Large Data Bases Journal (VLDB)*, 20(5):695–719, 2011. ISSN 1066-8888. Cité page 16.

F. Giannotti, M. Nanni, F. Pinelli et D. Pedreschi : Trajectory pattern mining. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 330–339, 2007. Cité pages 17 et 28.

F. Giannotti et D. Pedreschi, éditeurs. *Mobility, Data Mining and Privacy - Geographic Knowledge Discovery*. Springer, 2008. ISBN 978-3-540-75176-2. Cité pages 16 et 28.

M. Goodchild et J. Zhang : *Uncertainty in Geographical Information*. Research monographs in geographic information systems. Taylor & Francis, 2002. ISBN 9-780-20347-132-6. Cité page 36.

J. Gudmundsson, M. Kreveld et B. Speckmann : Efficient detection of motion patterns in spatio-temporal data sets. *In 13th International Symposium of ACM Geographic Information Systems*, pages 250–257, 2004. Cité pages 17 et 28.

P.N. Hai, D. Ienco, P. Poncelet et M. Teisseire : Mining representative movement patterns through compression. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, PAKDD'13, pages 314–326, 2013. Cité pages 18 et 28.

P.N. Hai, P. Poncelet et M. Teisseire : Get_move : an efficient and unifying spatio-temporal pattern mining algorithm for moving objects. *In 11th international conference on Advances in Intelligent Data Analysis*, IDA'12, pages 276–288, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-34155-7. Cité pages 18 et 28.

J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal et M.C. Hsu : Freespan : frequent pattern-projected sequential pattern mining. *In ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD'00, pages 355–359, New York, NY, USA, 2000. ACM. ISBN 1-58113-233-6. Cité page 53.

T. Ho, T.D. Nguyen et D.D. Nguyen : Visualization support for a user-centered KDD process. *In 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD'02, pages 519–524, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X. Cité page 107.

Y. Huang, S. Shekhar et H. Xiong : Discovering colocation patterns from spatial data sets : a general approach. *IEEE Transactions on Knowledge and Data Engineering TKDE*, 16 (12):pages 1472–1485, décembre 2004. ISSN 1041-4347. Cité pages 19 et 28.

Y. Huang, L. Zhang et P. Zhang : A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering TKDE*, 20(4):433–448, 2008. ISSN 1041-4347. Cité pages 22, 25 et 28.

F. Hussain, H. Liu, E. Suzuki et H. Lu : Exception rule mining with a relative interestingness measure. *In Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, PADKK'00, pages 86–97, London, UK, UK, 2000. Springer-Verlag. ISBN 3-540-67382-2. Cité page 61.

A. Inokuchi, T. Washio et H. Motoda : An apriori-based algorithm for mining frequent substructures from graph data. *In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, ECML/PKDD'00, pages 13–23, 2000. Cité page 4.

M. Jalali-Heravi et O.R. Zaïane : A study on interestingness measures for associative classifiers. *In ACM Symposium on Applied Computing*, SAC'10, pages 1039–1046, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-639-7. Cité page 62.

R. Johnston : *Order in spacegeography as a discipline in distance*. Johnston R, Williams M. Oxford, Oxford University Press, 2003. Cité page 29.

M. V. Joshi, G. Karypisx et V. Kumar : A Universal Formulation of Sequential Patterns. *In KDD 2001 workshop on Temporal Data Mining*, volume 1, page 7, 2001. Cité page 44.

A. Julea, N. Meger et P.H. Bolon : On mining pixel based evolution classes in satellite image time series. *In 5th Conference on Image Information Mining : pursuing automation of geospatial intelligence for environment and security*, ESA-EUSC'08, page 6, 2008. Cité page 39.

M. Kaufmann et D. Wagner : *Drawing Graphs : Methods and Models*, volume 2025 de *Lecture Notes in Computer Science*. Springer, 2001. ISBN 9783540420620. Cité page 110.

D.A. Keim : Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002. ISSN 1077-2626. Cité page 107.

K. Koperski et J. Han : Discovery of spatial association rules in geographic information databases. *In 4th International Symposium on Advances in Spatial Databases*, SSD'95, pages 47–66, 1995. ISBN 3-540-60159-7. Cité pages 18, 19, 28 et 34.

H.C.M. Kum, S. Paulsen et W. Wang : Comparative study of sequential pattern mining models. *In Foundations of Data Mining and knowledge Discovery*, volume 6 de *Studies in Computational Intelligence*, pages 43–70. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-26257-2. Cité page 59.

I.C. Lerman et J. Azé : A new probabilistic measure of interestingness for association rules, based on the likelihood of the link. *In Quality Measures in Data Mining*, volume 43 de *Studies in Computational Intelligence*, pages 207–236. Springer, 2007. ISBN 978-3-540-44911-9. Cité page 61.

Z. Li, J. Han, M. Ji, L.A. Tang, Y. Yu, B. Ding, J.G. Lee et R. Kays : Movemine : Mining moving object data for discovery of animal movement patterns. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(4):37 :1–37 :32, 2011. ISSN 2157-6904. Cité page 16.

J. Lin et Y. Li : Finding structural similarity in time series data using bag-of-patterns representation. *In International Conference on Scientific and Statistical Database Management*, SSDBM'09, pages 461–477, 2009. Cité pages 19 et 28.

F.A. Lisi et D. Malerba : Inducing multi-level association rules from multiple relations. *Machine Learning*, 55(2):175–210, 2004. Cité pages 18 et 28.

L. Mabit, N. Selmaoui-Folcher et F. Flouvat : Modélisation de la dynamique de phénomènes spatio-temporels par des séquences de motifs. *In Extraction et gestion des connaissances*, volume 20 de *EGC'11*, pages 455–466, 2011. ISBN 978-2-70568-112-8. Cité pages 22 et 28.

N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao et D.W. Cheung : Mining, indexing and querying historical spatiotemporal data. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 236, 2004. Cité pages 16 et 28.

H. Mannila, H. Toivonen et A.I. Verkamo : Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289, 1997. Cité page 4.

A. Marascu et F. Masseglia : Mining sequential patterns from data streams : a centroid approach. *Journal of Intelligent Information Systems*, 27(3):291–307, 2006. Cité page 39.

F. Masseglia, F. Cathala et P. Poncelet : The PSP approach for mining sequential patterns. *Principles of Data Mining and Knowledge Discovery*, pages 176–184, 1998. Cité pages 4 et 53.

F. Masseglia, P. Poncelet, M. Teisseire et A. Marascu : Web usage mining : extracting unexpected periods from web logs. *Data Mining and Knowledge Discovery (DMKD)*, 16(1):39–65, 2008. Cité page 39.

R. Mazza : *Introduction to Information Visualization*. Springer Publishing Company, Incorporated, 1 édition, 2009. ISBN 1848002181, 9781848002180. Cité page 107.

K. McGarry : A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20(1):39–61, mars 2005. ISSN 0269-8889. Cité page 44.

J. Mennis et J.W. Liu : Mining association rules in spatio-temporal data : An analysis of urban socioeconomic and land cover change. *Transactions in GIS*, 9(1):5–17, 2005. ISSN 1467-9671. Cité page 33.

P. Mohan, S. S., J.A. Shine et J.P. Rogers : Cascading spatio-temporal pattern discovery. *Knowledge and Data Engineering, IEEE Transactions on*, 24(11):1977–1992, 2012. ISSN 1041-4347. Cité pages 22, 26, 28, 44, 46 et 57.

B. Mortazavi-Asl, H. Pinto et U. Dayal : PrefixSpan : mining sequential patterns efficiently by prefix-projected pattern growth. *17th International Conference on Data Engineering*, pages 215–224, 2000. Cité pages 25, 33, 39, 53 et 82.

S. Nadi et M.R. Delavar : Spatio-temporal modeling of dynamic phenomena in GIS. *In ScanGIS*, pages 215–225, 2003. Cité page 5.

M. Nanni et D. Pedreschi : Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27(3):267–289, 2006. ISSN 0925-9902. Cité pages 17 et 28.

C. Pasquier, J. Sanhes, F. Flouvat et N Selmaoui-Folcher : Frequent pattern mining in attributed trees. *In Advances in Knowledge Discovery and Data Mining*, PAKDD'13, pages 26–37. Springer Berlin Heidelberg. Cité page 125.

J. Pei, J. Han, B. Mortazavi-Asl et H. Zhu : Mining access patterns efficiently from web logs. *In Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference*, PADKK'00, pages 396–407. Springer, 2000. Cité page 39.

D. Perera, J. Kay, I. Koprinska, K. Yacef et O.R. Zaïane : Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21:759–772, 2009. ISSN 1041-4347. Cité page 39.

D.J. Peuquet : It's about time : A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441–461, 1994. ISSN 00045608. Cité page 109.

Y. Ping, T. Xinming et W. Shengxiao : Dynamic cartographic representation of spatio-temporal data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVII(XXXVII-B2):7–12, 2008. Cité page 108.

S. Prinke, M. Wojciechowski et M. Zakrzewicz : Pruning discovered sequential patterns using minimum improvement threshold. *Foundations of Computing and Decision Sciences*, Vol. 31, No. 1:43–57, 2006. Cité page 62.

F. Qi et A.X. Zhu : Knowledge discovery from soil maps using inductive learning. *International Journal of Geographical Information Science*, pages 771–795, 2003. Cité page 33.

F. Qian, Q. He et J.F. He : Mining Spread Patterns of Spatio-temporal Co-occurrences over Zones. *Computational Science and Its Applications (ICCSA'09)*, pages 677–692, 2009. Cité pages 21 et 28.

J. Rabatel, S. Bringay et P. Poncelet : Aide à la décision pour la maintenance ferroviaire préventive. *In Extraction et Gestion des Connaissances*, EGC'10, pages 363–368. Cépaduès-Éditions, 2010. Cité page 39.

M.J.S. Rudwick : *The great Devonian controversy : the shaping of scientific knowledge among gentlemanly specialists / Martin J.S. Rudwick*. University of Chicago Press, Chicago :, 1985. ISBN 0226731014. Cité page 2.

A. Sallaberry, N. Pecheur, S. Bringay, M. Roche et M. Teisseire : Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. *Journal of Biomedical Informatics*, 44(5):760–774, 2011. ISSN 1532-0464. Cité page 107.

P. Salle, S. Bringay et M. Teisseire : Mining discriminant sequential patterns for aging brain. *In Artificial Intelligence in Medicine*, volume 5651 de *Lecture Notes in Computer Science*, pages 365–369. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-02975-2. Cité page 39.

H. Saneifar, S. Bringay, A. Laurent et M. Teisseire : S2MP : Similarity measure for sequential patterns. *In AusDM*, pages 95–104. Australian Computer Society, 2008. ISBN 978-1-920682-68-2. Cité pages 62 et 86.

J. Sanhes, F. Flouvat, C. Pasquier, N. Selmaoui-Folcher et J.F. Boulicaut : Extraction de motifs condensés dans un unique graphe orienté acyclique attribué. *In Extraction et Gestion des Connaissances*, RNTI E-24 EGC'13, pages 205–216. Hermann-Editions, 2013. Cité page 125.

N. Selmaoui-Folcher et F. Flouvat : How to use "classical" tree mining algorithms to find complex spatio-temporal patterns ? *In Database and Expert Systems Applications - 22nd International Conference*, volume 6861 de *DEXA'11*, pages 107–117, 2011. Cité pages 22 et 28.

N. Selmaoui-Folcher, F. Flouvat et D. Gay : Visualisation des motifs spatiaux. *Revue des Nouvelles Technologies de l'Information RNTI*, A-4:35–57, 2010. Cité pages 19 et 28.

N. Selmaoui-Folcher, F. Flouvat, D. Gay et I. Rouet : Spatial pattern mining for soil erosion characterization. *International Journal of Agricultural and Environmental Information Systems IJAEIS*, 2(2):73–92, 2011. Cité pages 19 et 28.

C. Sengstock, M. Gertz et Tran Van Canh : Spatial interestingness measures for co-location pattern mining. *In Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 821–826, 2012. Cité page 62.

S. Shekhar : A joinless approach for mining spatial colocation patterns. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1323–1337, 2006. ISSN 1041-4347. Cité pages 19 et 28.

S. Shekhar et Y. Huang : Discovering spatial co-location patterns : A summary of results. *In* Christian S. Jensen, Markus Schneider, Bernhard Seeger et Vassilis J. Tsotras, éditeurs : *SSTD*, volume 2121 de *Lecture Notes in Computer Science*, pages 236–256. Springer, 2001. ISBN 3-540-42301-X. Cité pages 18, 19, 26, 28 et 106.

A. Silberschatz et A. Tuzhilin : What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996. ISSN 1041-4347. Cité page 60.

R.T. Snodgrass : Temporal databases. *In* A.U. Frank, I. Campari et U. Formentini, éditeurs : *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, volume 639 de *Lecture Notes in Computer Science*, pages 22–64. Springer Berlin Heidelberg, 1992. ISBN 978-3-540-55966-5. Cité page 2.

R. Srikant et R. Agrawal : Mining sequential patterns : Generalizations and performance improvements. *In 5th International Conference on Extending Database Technology : Advances in Database Technology*, EDBT'96, pages 3–17, London, UK, UK, 1996. Springer-Verlag. ISBN 3-540-61057-X. Cité page 49.

I. Subasic et B. Berendt : Web mining for understanding stories through graph visualisation. *In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 570–579, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3502-9. Cité page 107.

P.N. Tan, V. Kumar et J. Srivastava : Selecting the right interestingness measure for association patterns. *In Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD'02, pages 32–41, 2002. Cité page 61.

A. Termier, M.C. Rousset et M. Sebag : Treefinder : a first step towards xml data mining. *In IEEE International Conference on Data Mining ICDM*, pages 450–457, 2002. Cité page 4.

I.I. Tsoukatos et D. Gunopulos : Efficient mining of spatiotemporal patterns. *Advances in Spatial and Temporal Databases*, pages 425–442, 2001. Cité pages 22, 27, 28, 34, 40, 102 et 106.

E.R. Tufte : *The visual display of quantitative information.* Numéro vol. 914 *in* The Visual Display of Quantitative Information. Graphics Press, 1983. Cité page 110.

J. Wang, W. Hsu et M.L. Lee : Mining generalized spatio-temporal patterns. *In Database Systems for Advanced Applications*, pages 649–661. Springer, 2005. Cité pages 22, 25 et 28.

J. Wang, W. Hsu, M.L. Lee et J. Wang : FlowMiner : finding flow patterns in spatio-temporal databases. *16th IEEE International Conference on Tools with Artificial Intelligence*, (Ictai): 14–21, 2004a. Cité pages 22, 23 et 28.

K. Wang, Y. Xu et J.X. Yu : Scalable sequential pattern mining for biological sequences. *In CIKM '04 : Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*, pages 178–187, New York, NY, USA, 2004b. ACM. ISBN 1-58113-874-1. Cité page 39.

L. Wang, L. Zhou, J. Lu et J. Yip : An order-clique-based approach for mining maximal co-locations. *Information Sciences*, 179(19):3370–3382, 2009. ISSN 0020-0255. Cité pages 19 et 28.

M. Ward, G.S. Grinstein et D. Keim : *Interactive Data Visualization : Foundations, Techniques, and Applications.* A. K. Peters, Ltd., Natick, MA, USA, 2010. ISBN 1568814739, 9781568814735. Cité page 107.

C. Ware : *Information Visualization : Perception for Design*. Interactive Technologies. Elsevier Science, 2004. ISBN 9780080478494. Cité page 111.

P.C. Wong, W. Cowley, H. Foote, E. Jurrus et J. Thomas : Visualizing sequential patterns for text mining. *In IEEE Symposium on Information Vizualization 2000*, INFOVIS'00, page 105, Washington, DC, USA, 2000. IEEE Computer Society. ISBN 0-7695-0804-9. Cité page 107.

T. Wu, Y. Chen et J. Han : Re-examination of interestingness measures in pattern mining : a unified framework. *Data Mining and Knowledge Discovery*, 21(3):371–397, 2010. ISSN 1384-5810. Cité page 44.

H. Yang, S. Parthasarathy et S. Mehta : A generalized framework for mining spatio-temporal patterns in scientific data. *In 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD'05, pages 716–721, New York, NY, USA, 2005. ACM. ISBN 1-59593-135-X. Cité pages 22 et 28.

J.S. Yoo et M. Bow : Mining top-k closed co-location patterns. *In Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on*, pages 100–105, 2011. Cité page 62.

M. Yuan : Knowledge toward discovery about geographic dynamics in spatiotemporal databases. *In Geographic Data Mining and Knowledge Discovery, Second Edition*, pages 347–365. Edited by Harvey J . Miller and Jiawei Han, 2009. Cité pages 4 et 11.

M.J. Zaki : Spade : An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60, 2001. Cité page 49.

M.J. Zaki : Efficiently mining frequent trees in a forest. *In 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 71–80. ACM press, 2002. Cité page 4.

Mohammed J. Zaki et Ching-Jui Hsiao : Charm : An efficient algorithm for closed association rule mining. Rapport technique, Computer science, Rensselaer Polytechnic Institute, 1999. Cité page 61.

K. Zeitouni, L. Yeh et M.A. Aufaure : Join indices as a tool for spatial data mining. *In Temporal, Spatial, and Spatio-Temporal Data Mining, First International Workshop*, volume 2007 de *TSDM'00*, pages 105–116. Springer, 2000. Cité pages 18 et 28.

Y. Zhao, H. Zhang, L. Cao, C. Zhang et H. Bohlscheid : Combined pattern mining : From learned rules to actionable knowledge. *In 21st Australasian Joint Conference on Artificial Intelligence : Advances in Artificial Intelligence*, AI'08, pages 393–403, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-89377-6. Cité page 61.